PhD Dissertation

International Doctorate School in Information and
Communication Technologies

DISI - University of Trento

Advanced Techniques for the Classification of Very High Resolution and Hyperspectral Remote Sensing Images

Claudio Persello

Advisor:
Prof. Lorenzo Bruzzone
Università degli Studi di Trento

February 2010

A mamma Liliana

Abstract

This thesis is about the classification of the last generation of very high resolution (VHR) and hyperspectral remote sensing (RS) images, which are capable to acquire images characterized by very high resolution from satellite and airborne platforms. In particular, these systems can acquire VHR multispectral images characterized by a geometric resolution in the order or smaller than one meter, and hyperspectral images, characterized by hundreds of bands associated to narrow spectral channels. This type of data allows to precisely characterizing the different materials on the ground and/or the geometrical properties of the different objects (e.g., buildings, streets, agriculture fields, etc.) in the scene under investigation. This remote sensed data provide very useful information for several applications related to the monitoring of the natural environment and of human structures. However, in order to develop real-world applications with VHR and hyperspectral data, it is necessary to define automatic techniques for an efficient and effective analysis of the data. Here, we focus our attention on RS image classification, which is at the basis of most of the applications related to environmental monitoring. Image classification is devoted to translate the features that represent the information present in the data in thematic maps of the land cover types according to the solution of a pattern recognition problem. However, the huge amount of data associated with VHR and hyperspectral RS images makes the classification problem very complex and the available techniques are still inadequate to analyze these kinds of data. For this reason, the general objective of this thesis is to develop novel techniques for the analysis and the classification of VHR and hyperspectral images, in order to improve the capability to automatically extract useful information captured from these data and to exploit it in real applications. Moreover we addressed the classification of RS images in operational conditions where the available reference labeled samples are few and/or not completely reliable (which is quite common in many real problems). In particular, the following specific issues are considered in this work:

1. *development of feature selection for the classification of hyperspectral images, for identifying a subset of the original features that exhibits at the same time high capability to discriminate among the considered classes and high invariance in the spatial domain of the scene;*
2. *classification of RS images when the available training set is not fully reliable, i.e., some labeled samples may be associated to the wrong information class (mislabeled patterns);*
3. *active learning techniques for interactive classification of RS images.*

4. *definition of a protocol for accuracy assessment in the classification of VHR images that is based on the analysis of both thematic and geometric accuracy;*

For each considered topic an in deep study of the literature is carried out and the limitations of currently published methodologies are highlighted. Starting from this analysis, novel solutions are theoretically developed, implemented and applied to real RS data in order to verify their effectiveness. The obtained experimental results confirm the effectiveness of all the proposed techniques.

Keywords:
Very high resolution images, hyperspectral images, supervised classification, semisupervised classification, support vector machines, context-based classification, feature selection, stationary features, noisy training sets, active learning, map accuracy assessment, thematic accuracy, geometric accuracy, remote sensing.

Ringraziamenti

Il contenuto di questa tesi è il risultato dell'attività di ricerca sviluppata in tre anni presso il laboratorio di telerilevamento dell'Università di Trento. Ringrazio sinceramente Lorenzo per avermi dato la possibilità di fare il dottorato, e per avermi seguito nella mia attività di ricerca dandomi continuamente idee sempre nuove ed interessanti spunti per l'approfondimento. Le nostre discussioni scientifiche sono sempre risultate un grande stimolo per la mia curiosità, la mia voglia di conoscere, migliorare e scoprire cose nuove. Queste mi hanno aiutato a credere nel lavoro che ho fatto anche quando i risultati erano scoraggianti e di raggiungere i risultati scientifici che sono riportati in questa tesi.

Ringrazio di cuore la mamma, che in questi anni di studio mi ha sempre sostenuto moralmente e spiritualmente! Con lei ho potuto condividere successi e momenti difficili o di incertezza. E anche di fronte a decisioni non facili ha sempre cercato che scegliessi la cosa migliore per me. Grazie per tutte quelle volte che il venerdì sera mi hai aspettato fino a tardi perché potessi raccontarti come era andata la settimana! Ringrazio Alessandro e Francesca che ho sempre sentito vicino in questi anni di studio.

Un grazie va anche a tutti i colleghi ed amici dell'RSLab con cui ho condiviso molto in questi tre anni di studio: Francesca, Michele, Dominik, Silvia, Mauro, Adamo, Luca, Michele, Swarnajyoti. Un caro pensiero va a Begüm con cui ho avuto la fortuna di lavorare e condividere momenti piacevoli durante il suo periodo di stage a Trento.

Acknowledgments

The content of this thesis is the result of the research activity developed in three years at the Remote Sensing Laboratory of the University of Trento. I am sincerely grateful to Lorenzo, who gave me the opportunity to attend the Ph.D. school, and guided me in my research activity giving constantly to me new ideas and hints about my research topics. Our scientific discussions were always very interesting for me and stimulated my curiosity, my wish to learn, to improve and discover new things. These discussions helped me to believe in my work even when the results were not good and helped me to reach the scientific results reported in this document.

I would like to thank mum, who always supported me morally and spiritually in all these years of study! I could share with her my achievements as well as the difficult and uncertain moments. She supported me in all important decisions wishing the best for me. Thank you for all the times that you waited for me on Friday night until I came home so that I could tell you how I spent the week! Thanks to Alessandro and Francesca that I always felt near me in all these years of study.

A particular thank goes to all my colleagues and friends of RSLab: Francesca, Michele, Dominik, Silvia, Mauro, Adamo, Luca, Michele, Swarnajyoti. With them I shared a lot in these last years. A special thank goes to Begüm, with whom I had the pleasure to work with and to spend nice moments during her internship in Trento.

Contents

Chapter 1

1. Introduction

In this chapter we introduce this dissertation presenting the background on remote sensing and a review on the last generation of remote sensing sensors characterized by very high geometrical and/or spectral resolution and their applications to environmental monitoring. We also describe the most critical issues related to the automatic analysis and classification of the data collected by these sensors, as well as the general motivations and objectives of this work. Furthermore, we present the specific issues taken into account in this research activity and the novel contributions of the thesis. Finally, the structure and organization of this document is described.

1.1 Overview on remote sensing systems

With the words "Remote Sensing" (RS) we refer to a technology capable to collect and to interpret information regarding an object without being directly in contact with the item under investigation. In particular, in this dissertation we take into account the use of RS images collected by sensors on board on aircrafts or spacecraft platforms for observing and characterizing the Earth surface. These sensors acquire the energy emitted and reflected from the Earth's surface to construct an image of the landscape beneath the platform [1]. Depending on the source of the energy involved in the image acquisition, two main kinds of RS imaging systems can be distinguished: 1) passive systems and 2) active systems.

Passive (or optical) systems rely on the presence of an external illumination source, i.e., the sun. The signal measured by the sensor is: 1) the radiation coming from the sun, that is reflected by the Earth surface and passing through the atmosphere arrives to the sensor; and 2) the energy emitted by the Earth itself, because of its own temperature. The energy measured by the sensor is usually collected in several spectral bands (the spectral range of each single band defines the spectral resolution) and over a certain elementary area (that defines the geometric resolution). After that, the measure is converted into an opportune electric signal and recorded as digital image. These sensors are usually called multispectral scanners. Sensors capable to collect the radiation in hundreds of very narrow spectral bands are called hyperspectral.

On the contrary, in active RS systems, the sensor itself (e.g., an antenna) emits the energy (an electromagnetic radiation) directed towards the Earth's surface and measures the energy scattered back to it. Radar systems, such as real aperture (RAR), synthetic aperture radar (SAR), and LIDAR are examples of active sensors. In these systems, the time delay between emission and return is measured to establish the location and height of objects, and the power of the received radiation provide information for characterizing the object under investigation.

In this dissertation, we focus on the analysis of optical multispectral and hyperspectral images and in particular on the last generation of sensors, which can provide images characterized by very high geometrical/spectral resolution.

1.2 Overview on the last generation of remote sensing imaging systems

In the last decade, the advances in imaging sensors and satellite technologies resulted in the development of a new generation of systems capable to acquire images characterized by very high resolution from satellite and airborne platforms. In particular, these systems can acquire: 1) very high resolution (VHR) multispectral images characterized by a geometric resolution in the order of (or smaller than) 1 m; and 2) hyperspectral images, characterized by hundreds of bands associated to narrow spectral channels. In the following subsections we will briefly review the last sensor advances in the field of VHR and hyperspectral imaging systems, respectively.

1.2.1 VHR satellites imaging systems

VHR images became available (and popular) with the launch of commercial satellites like Ikonos and Quickbird, with on-board multispectral scanners characterized by a geometrical resolution in the order of 1 m. These satellites can acquire four multispectral bands, in the visible and near infrared spectral ranges, and a panchromatic channel with four time higher spatial resolution. These satellites represent a significant improvement in the geometric resolution with respect to the popular Landsat satellites. Indeed, Landsat 7 (the last satellite of the Landsat program) provides seven multispectral bands in the visible, near and thermal infrared ranges with a geometric resolution of 30 m (except the thermal infrared band that has a resolution of 60 m) and a panchromatic channel with a spatial resolution of 15 m. The SPOT 5 satellite, the last launched and operating satellite of the SPOT program, can acquire four multispectral bands in the ranges of visible, near, and mid infrared with a spatial resolution of 10 m (except the mid infrared band that has a resolution of 20 m) and a panchromatic band with a maximum resolution of 2.5 m. Recently, a new generation of VHR satellite systems became available, i.e., GeoEye-1, WorldView-1 and 2, which further improve the geometric resolution, providing a panchromatic channel with a resolution smaller that half meter. It is interesting to note that the WorldView-2 satellite increase the spectral resolution other than the geometric resolution, by providing eight channels instead of the common four. Moreover, in the next years the quality and the availability of this type of data are going to further increase thanks to the missions GeoEye-2 and Pleiades. Table 1.1 reports the main characteristics of the most popular satellite systems of the last decade with on board multispectral scanners. Fig. 1.1 shows a graph of the increase of the spatial resolution of popular satellites with on board multispectral systems since 1970.

Table 1.1 – Main characteristics of multispectral sensors on board of satellite platforms. All the considered satellites are in a polar sun-synchronous orbit with equatorial crossing time 10 a.m.

Satellite	Sensor bands [nm]	Spatial resolution (at nadir)	Swath width	Orbit altitude	Year of launch
Landsat 7	520-900 (pan) 450-520 (blue) 520-600 (green) 630-690 (red) 760-900 (NIR) 1550-1750 (MIR 1) 10400-12500 (TIR) 2080-2350 (MIR 2)	15 m 30 m 30 m 30 m 30 m 30 m 60 m 30 m	185 km	705 km	1999
Ikonos	526-900 (pan) 445-516 (blue) 505-595 (green) 0.632-0.698 (red) 0.757-0.853 (NIR)	0.82 m 3.2 m	11 km	681 km	1999
Eros A	500-900 (pan)	1.8 m	14 km	480 km	2000
Quickbird	445-900 (pan) 450-520 (blue) 520-600 (green) 630-690 (red) 760-900 (NIR)	0.61 m 2.44 m	16.5 km	450 km	2001
SPOT 5	480-710 (pan) 500-590 (green) 610-680 (red) 780-890 (NIR) 1580-1750 (MIR)	2.5 m 10 m 10 m 10 m 20 m	60 km	832 km	2002
Eros B	500-900 (pan)	0.7 m	7 km	600 km	2006
WorldView 1	450-900 (pan)	0.50 m	17.6 km	496 km	2007
GeoEye 1	450-900 (pan) 450-520 (blue) 520-600 (green) 625-695 (red) 760-900 (NIR)	0.41 m 1.65 m	15.2 km	681 km	2008

Satellite	Sensor bands [nm]	Spatial resolution (at nadir)	Swath width	Orbit altitude	Year of launch
WorldView 2	450-800 (pan) 400-450 (coastal) 450-510 (blue) 510-580 (green) 585-625 (yellow) 630-690 (red) 705-745 (red edge) 770-895 (NIR 1) 860-1040 (NIR 2)	0.46 m 1.84 m	16.4 km	770 km	2009
Pleiades-HR 1	480-830 (pan) 430-550 (blue) 490-610 (green) 600-720 (red) 750-950 (NIR)	0.7 m 2.8 m	20 km	694 km	2010
Pleiades-HR 2	480-830 (pan) 430-550 (blue) 490-610 (green) 600-720 (red) 750-950 (NIR)	0.7 m 2.8 m	20 km	694 km	2011
GeoEye 2	Pan	0.25 m	-	-	2012

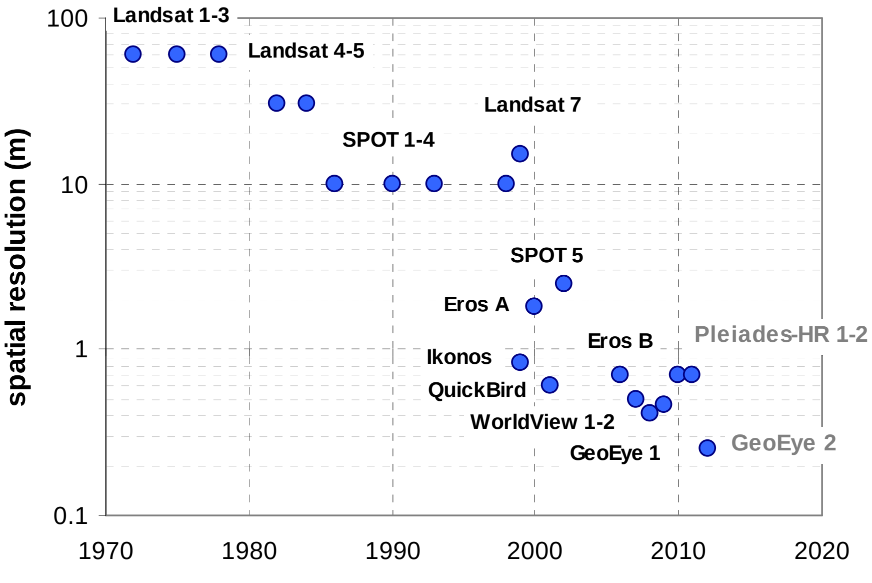


Fig. 1.1 – Spatial resolution of multispectral satellite sensors

VHR images allow one the precise recognition of the shape and the geometry of the objects present on the ground as well as the identification of the different land-cover classes. For these reasons, VHR data are very important sources of information for the development of many applications related to the monitoring of natural environments and human structures. Strategic applications for public administrations are related to the monitoring and the management of natural resources, agriculture fields, urban areas or for analyzing evacuation planning in areas with the risk of floods or fires. Other examples of interesting applications are building detection and building abuse discovering, road networks extraction and road map updating.

1.2.2 Hyperspectral imaging systems

Hyperspectral sensors can acquire hundreds of bands associated to narrow spectral channels, allowing a dense sampling of the spectral signature of the land-covers. At the present, the acquisition of hyperspectral images can be obtained by airborne platforms, or by MODIS, CHRIS/Proba, and Hyperion systems, which are the only satellites with on-board hyperspectral sensors that acquire images in some tens or hundreds of bands. Table 1.2 reports the main recent hyperspectral sensors and their spectral characteristics. However, among the others, in the next years the Italian Space Agency (ASI) and the German Space Agency (DLR) are going to launch two new satellite missions with high resolution hyperspectral sensors, called PRISMA and EnMAP, respectively. The PRISMA sensor will combine a hyperspectral sensor (operating in the spectral range 400-2500 nm with spectral resolution of 10 nm) that has a geometrical resolution of 20-30 m with a panchromatic camera capable to acquire images with a geometrical resolution of 2.5-5 m. This combination will allow one to precisely characterize both the different types of materials on the ground as well as the shape and the geometrical properties of the objects in the scene under investigation.

Hyperspectral images represent a very rich source of information for a precise recognition and characterization of the materials and objects on the ground. Hyperspectral images allow one the development of important applications like the detailed classification of forest areas, pollution monitoring, analysis of inland water and coastal zones, analysis of natural risks (fires, floods, eruptions, earthquakes), etc.

Table 1.2 - Recent hyperspectral sensors and related spectral properties [2].

Sensor name	**Manufacturer**	**Platform**	**Maximum Number of Bands**	**Maximum Spectral Resolution**	**Spectral range**
Hyperion on EO-1	NASA Goddard Space Flight Center	satellite	220	10 nm	0.4 – 2.5 µm
MODIS	NASA	satellite	36	40 nm	0.4 – 14.3 µm
CHRIS Proba	ESA	satellite	up to 63	1.25 nm	0.415 – 1.05 µm
AVIRIS	NASA Jet Propulsion Lab	aerial	224	10 nm	0.4 – 2.5 µm
HYDICE	Naval Research Lab	aerial	210	7.6 nm	0.4 – 2.5 µm
PROBE-1	Earth Search Sciences Inc.	aerial	128	12 nm	0.4 – 2.45 µm
CASI 550	ITRES Research Limited	aerial	288	1.9 nm	0.4 – 1 µm

Sensor name	Manufacturer	Platform	Maximum Number of Bands	Maximum Spectral Resolution	Spectral range
CASI 1500	ITRES Research Limited	aerial	288	2.5 nm	0.4 – 1.05 µm
SASI 600	ITRES Research Limited	aerial	100	15 nm	0.95 – 2.45 µm
TASI 600	ITRES Research Limited	aerial	64	250 nm	8 – 11.5 µm
HyMap	Integrated Spectronics	aerial	125	17 nm	0.4 – 2.5 µm
ROSIS	DLR	aerial	84	7.6 nm	0.43 – 0.85 µm
EPS-H (Environmental Protection System)	GER Corporation	aerial	133	0.67 nm	0.43 – 12.5 µm
EPS-A (Environmental Protection System)	GER Corporation	aerial	31	23 nm	0.43 – 12.5 µm
DAIS 7915 (Digital Airborne Imaging Spectrometer)	GER Corporation	aerial	79	15 nm	0.43 – 12.3 µm
AISA Eagle	Spectral Imaging	aerial	244	2.3 nm	0.4 to 0.97 µm
AISA Eaglet	Spectral Imaging	aerial	200	-	0.4 to 1.0 µm
AISA Hawk	Spectral Imaging	aerial	320	8.5 nm	0.97 to 2.45 µm
AISA Dual	Spectral Imaging	aerial	500	2.9 nm	0.4 to 2.45 µm
MIVIS (Multispectral Infrared and Visible Imaging Spectrometer)	Daedalus	aerial	102	20 nm	0.43 – 12.7 µm
AVNIR	OKSI	Aerial	60	10 nm	0.43 – 1.03 µm

1.3 Motivation, objectives and novel contributions of this thesis

In order to develop the applications mentioned in the previous sections with VHR and hyperspectral data, it is necessary to define automatic techniques for an efficient and effective analysis of the data. Here, we focus our attention on RS image classification, which is at the basis of the development of most of the aforesaid applications, and is devoted to translate the features that represent the information present in the data in thematic maps of the land cover types according to the solution of a pattern recognition problem. However, the huge amount of data associated with VHR and hyperspectral RS images makes the classification problem very complex. At the state of the art, the most promising techniques for the classification of the last generation of RS data are based on kernel methods and support vector machines [3]-[4], which revealed very effective and robust in the solution of many classification problems. Nonetheless, the available techniques are still inadequate to analyze these kinds of data and further investigations are needed to effectively exploit VHR and hyperspectral images for the development of real-world applications. For this reason, the general objective of this thesis is to develop novel techniques for the analysis and the classification of VHR and hyperspectral images, in order to improve the capability to automatically extract the useful information captured from these data and to exploit it in real applications. We addressed these problems also considering operational conditions where the available reference labeled samples are few and/or not completely reliable (which is quite common in many real problems). In particular, the following specific issues are considered in this work:

a) selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability;
b) classification of RS images when the available training set is not fully reliable, i.e., some labeled samples may be associated to the wrong information class (mislabeled patterns);
c) active learning techniques for interactive classification of RS images.
d) definition of a protocol for accuracy assessment in the classification of VHR images that is based on the analysis of both thematic and geometric accuracy;

In order to address the abovementioned issues, we developed novel approaches and techniques for the analysis and classification of RS images. The main goals of these methods are briefly described in the following.

a) A Novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability

Hyperspectral RS images, which are characterized by a dense sampling of the spectral signature of different land-cover types, represent a very rich source of information for the analysis and automatic recognition of land-cover classes. However, supervised classification of hyperspectral images is a very complex methodological problem due to many different issues: 1) the small value of the ratio between the number of training samples and the number of available spectral channels (and thus of classifier parameters), which results in the Hughes phenomenon [5]; 2) the high correlation among training patterns taken from the same area, which violates the required assumption of independence of samples included in the training set (thus reducing the information conveyed to the classification algorithm by the considered samples); and 3) the nonstationary behavior of the spectral signatures of land-cover classes in the spatial domain of the scene, which is due to physical factors related to ground (e.g., different soil moisture or composition), vegetation, and atmospheric conditions. All the aforementioned issues result in decreasing the robustness, the generalization capability, and the overall accuracy of classification systems used to generate the land-cover maps.

In this thesis, we address the aforementioned problem by proposing a novel approach to feature selection that, unlike standard techniques, aims at identifying a subset of features that exhibits both high discrimination ability among the considered classes and high invariance in the spatial domain of the investigated scene. This approach is implemented by defining a novel criterion function that is based on the evaluation of two terms: 1) a standard separability measure and 2) a novel invariance measure that assesses the stationarity of features in the spatial domain. The search algorithm, adopted for deriving the subsets of features that jointly optimize the two terms, is based on the optimization of a multiobjective problem for the estimation of the Pareto-optimal solutions [6]. For the assessment of the two terms of the criterion function, we propose both a supervised and a semisupervised method that can be alternatively adopted depending on the amount of available reference data and on their properties. The proposed approach can be integrated in the design of any system for hyperspectral image classification (e.g., based on parametric or distribution-free supervised algorithms) for increasing the robustness and the generalization capability of the classifier.

b) A novel context-sensitive semisupervised SVM classifier robust to mislabeled training samples

The classification of RS images is often performed by using supervised classification algorithms, which require the availability of labeled samples for the training of the classification model. All these algorithms are sharply affected from the quality of the labeled samples used for training the classifier, whose reliability is of fundamental importance for an adequate learning of the properties of the investigated scene (and, thus, for obtaining accurate classification maps). In supervised classification, the implicit assumption is that all labels associated with training patterns are correct. Unfortunately, in many real cases, this assumption does not hold, and small amounts of training samples are associated with a wrong information class due to errors occurred in the phase of collection of labeled samples. Labeled samples can be derived by the following: 1) *in situ* ground truth surveys; 2) analysis of reliable reference maps; or 3) image photointerpretation. In all these cases, mislabeling errors are possible. During the ground truth surveys, mislabeling errors may occur due to imprecise geolocalization of the positioning system; this leads to the association of the identified land-cover label with a wrong geographic coordinate and, thus, with the wrong pixel (or region of interest) in the remotely sensed image. Similar errors may occur if the image to be classified is not precisely georeferenced. When reference maps are used for extracting label information, possible errors present in the maps propagate to the training set. The case of image photointerpretation is also critical, as errors of the human operator may occur, leading to a mislabeling of the corresponding pixels or regions. Mislabeled patterns bring distort (wrong) information to the classifier. The effect of noisy patterns in the learning phase of a supervised classifier is to introduce a bias in the definition of the decision regions, thus decreasing the accuracy of the final classification map.

In this thesis, we address the aforementioned problems by the following: 1) presenting a novel context-sensitive semisupervised support vector machine (CS^{4}VM) classification algorithm, which is robust to noisy training sets, and 2) analyzing the effect of noisy training patterns and of their distribution on the classification accuracy of widely used supervised and semisupervised classifiers. The main idea behind the proposed methodology is to exploit the information of the context patterns to reduce the bias effect of the mislabeled training patterns on the definition of the discriminating hyperplane of the SVM classifier, thus decreasing the sensitivity of the learning algorithm to unreliable training samples. This is accomplished by explicitly including the samples belonging to the neighborhood system of each training pattern in the definition of the cost function used for the learning of the classifier. These samples are considered by exploiting the labels derived through a semisupervised classification process (for this reason, they are called semilabeled samples). The semilabeled context patterns have the effect to mitigate the bias introduced by noisy patterns by adjusting the position of the hyperplane. This strategy is defined according to a learning procedure for the proposed CS^{4}VM that is based on two main steps: 1) supervised learning with original training samples and classification of the (unlabeled) context patterns and 2) contextual semisupervised learning based on both original labeled patterns and semilabeled context patterns according to a novel cost function.

c) Batch mode active learning methods for interactive classification of RS images

As mentioned before, automatic classification of RS images is generally performed by using supervised classification techniques, which require the availability of labeled samples for training the supervised algorithm. The amount and the quality of the available training samples are crucial for obtaining accurate classification maps. However, in many real world problems the available training samples are not enough for an adequate learning of the classifier. In order to enrich the information given as input to the learning algorithm (and to improve classification accuracy) semisupervised approaches can be adopted to jointly exploit labeled and unlabeled samples in the training of the classifier. Semisupervised approaches based on Support Vector Machines (SVMs) have been successfully applied to the classification of multispectral and hyperspectral data, where the ratio between the number of training samples and the number of available spectral channels is small. However, an alternative and conceptually different approach to improve the statistic in the learning of a classifier is to iteratively expand the original training set according to an interactive process that involves a supervisor. This approach is known in the machine learning community as active learning [7]-[9], and although marginally considered in the RS community, can result very useful in different application domains. In active learning: 1) the learning process repeatedly queries available unlabeled samples to select the ones that are expected to be the most informative for an effective learning of the classifier, 2) the supervisor (e.g., the user) labels the selected samples interacting with the system, and 3) the learner updates the classification rule by retraining with the updated training set. Therefore, the unnecessary and redundant labeling of non informative samples is avoided, greatly reducing the labeling cost and time. Moreover, active learning allows one to reduce the computational complexity of the training phase.

In this thesis we investigate different batch mode AL techniques proposed in the machine learning literature and we properly generalize them to the classification of RS images with multiclass problem addressed by support vector machines (SVMs). The key issue of batch mode AL is to select sets of samples with little redundancy, so that they can provide the highest possible information to the classifier. Thus, the query function adopted for selecting the batch of the most informative samples should take into account two main criteria: 1) uncertainty, and 2) diversity of samples. The uncertainty criterion is associated to the confidence of the supervised algorithm in correctly classifying the considered sample, while the diversity criterion aims at selecting a set of unlabeled samples that are as more diverse (distant one another) as possible, thus reducing the redundancy among the selected samples. The combination of the two criteria results in the selection of the potentially most informative set of samples at each iteration of the AL process. Moreover, we propose a novel query function that is based on a kernel clustering technique for assessing the diversity of samples and a new strategy for selecting the most informative representative sample from each cluster. The investigated and proposed techniques are theoretically and experimentally compared among them and with other AL algorithms proposed in the RS literature in the classification of VHR images and hyperspectral data. On the basis of this comparison some guidelines are derived on the use of AL techniques for the classification of different types of RS images.

d) A novel protocol for accuracy assessment in classification of very high resolution images

With the availability of VHR images acquired by satellite multispectral scanners, it is possible to acquire detailed information on the shape and the geometry of the objects present on the ground. This detailed information can be exploited by automatic classification systems to generate land-cover maps that exhibit a high degree of geometrical details. The precision that the classification system can afford in the characterization of the geometrical properties of the objects present on the ground is particularly relevant in many practical applications, e.g., in urban area mapping, building characterization, target detection, crop fields classification in precision farming, etc. In this context, a major open issue in classification of VHR images is the lack of adequate strategies for a precise evaluation of the quality of the produced thematic maps. The most common accuracy assessment methodology in classification of VHR images is based on the computation of thematic accuracy measures according to collected reference data. However, the thematic accuracy alone does not result sufficient for effectively characterizing the geometrical properties of the objects recognized in a map, because it assesses the correctness of the land-cover labels of sparse test pixels (or regions of interests) that do not model the actual shape of the objects in the scene. Thus, often maps derived by different classifiers (or with different parameter values for the same classifier) that have similar thematic accuracy exhibit significantly different geometric properties (and thus global quality). For this reason, in many real classification problems the quality of the maps obtained by the classification of VHR data is assessed also through a visual inspection. However, this procedure can provide just a subjective evaluation of the map quality that can not be quantified. Thus, it is important to develop accuracy assessment protocols for a precise, objective, and quantitative characterization of the quality of thematic maps in terms of both thematic and geometric properties. These protocols could be used not only for assessing the quality of thematic maps generated by different classification systems, but also for better driving the model selection of a single classifier, i.e., the selection of the optimum values for the free parameters of a supervised categorization algorithm.

Here, we address the abovementioned problem by proposing a novel protocol for a precise, automatic, and objective characterization of the accuracy of thematic maps derived from VHR images. The proposed protocol is based on the evaluation of two families of indices: 1) thematic accuracy indices; and 2) a set of novel geometric indices that assess different properties of the objects recognized in the thematic map. The proposed protocol can be used to: 1) objectively characterize the thematic and geometric properties of classification maps; 2) to select the map that better fit specific user requirements; or 3) to identify the map that exhibits in average best global properties if no specific requirements are defined. Moreover, we propose a novel approach for tuning the free parameters of supervised classification algorithms (e.g., SVM), which is based on the optimization of a multiobjective problem. The aim of this approach is to select the parameter values that result in a classification map that exhibits high geometric and thematic accuracies.

1.4 Structure of the Thesis

This thesis is organized in seven chapters. The present chapter presented a brief overview on both RS and the last generation of VHR and hyperspectral sensors. In addition, it introduced the background, the motivation and the main novel contributions of this thesis. The rest of the chapters are aimed at addressing the issues introduced in section 1.3, by presenting the analysis of the state of the art and proposing the novel techniques and approaches developed during the Ph.D. activity.

Chapter 2 presents an extensive and critical review on the use of kernel methods and in particular of support vector machines (SVMs) in the classification of RS data.

Chapter 3 proposes a novel approach to feature selection for the classification of hyperspectral images that aims at selecting a subset of the original features that exhibits at the same time high capability to discriminate among the considered classes and high invariance in the spatial domain of the scene. This approach results in a more robust classification system with improved generalization properties with respect to standard feature-selection methods.

Chapter 4 describes a novel context-sensitive semisupervised support vector machines (CS^{4}VM) classifier, which is aimed at addressing classification problems where the available training set is not fully reliable, i.e., some labeled samples may be associated to the wrong information class.

Chapter 5 presents an analysis on the use of active learning techniques for the interactive classification of RS images and a comparison of active learning techniques based on SVM generalized to multiclass problems. Moreover, a novel query function for the selection of a batch of unlabeled samples to be included in the training set is proposed.

Chapter 6 introduces a novel protocol for the accuracy assessment of the thematic maps obtained by the classification of VHR images. The proposed protocol is based on the analysis of two families of indices: 1) the traditional thematic accuracy indices; and 2) a set of novel geometric indices that model different geometric properties of the objects recognized in the map.

Chapter 7 draws the conclusions of this thesis. Furthermore, future developments of the research activity are discussed.

1.5 Reference

[1] J.A. Richards, X. Jia, “Remote sensing digital image analysis”, 4th ed., Springer-Verlag, New York, 2006.

[2] M. Dalponte, L. Bruzzone, L. Vescovo, D. Giannelle, “The role of spectral resolution and classifier complexity in the analysis of hyperspectral images of forest areas”, *Remote Sensing of the Environment*, vol. 113, no. 11, pp. 2345-2355, 2009.

[3] V. N. Vapnik, The Nature of Statistical Learning Theory, 2nd ed., New York: Springer, 2001.

[4] F. Melgani, L. Bruzzone, “Classification of Hyperspectral Remote Sensing Images With Support Vector Machines,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778- 1790, Aug. 2004.

[5] G. F. Hughes, “On the mean accuracy of statistical pattern recognition,” *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 55–63, January 1968.

[6] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II”, *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182-197, 2002.

[7] M. Li and I. Sethi, “Confidence-Based active learning”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1251-1261, 2006.

[8] D. D. Lewis and W. A. Gale, “A sequential algorithm for training text classifiers,” in Proc. 17th Annu. Int. ACM-SIGIR Conf. Res. Dev. Inf. Retrieval, W. B. Croft and C. J. van Rijsbergen, Eds., London, U.K., pp. 3–12, 1994.

[9] K. Brinker, “Incorporating Diversity in Active Learning with Support Vector Machines,” *Proceedings of the International Conference on Machine Learning*, Washington DC, pp. 59-66, 2003.

Chapter 2

2. Support Vector Machine for the Classification of Remote Sensing Data

This chapter presents an extensive and critical review on the use of kernel methods and in particular of support vector machines (SVMs) for the classification of remote sensing (RS) data. The chapter recalls the mathematical formulation and the main theoretical concepts related to SVMs, and discusses the motivations at the basis of the use of SVMs in remote sensing. A review on the main applications of SVMs in classification of remote sensing is given, presenting a literature survey on the use of SVMs for the analysis of different kinds of RS images. In addition, the most recent methodological developments related to SVM-based classification techniques in RS are illustrated by focusing on semisupervised, domain adaptation, and context-sensitive approaches. Finally, the most promising research directions on SVM in RS are identified and discussed

2.1 Introduction

In the last two decades there have been significant improvements both in the technology associated with the development of the sensors used in RS to acquire signals and images for Earth observation (as reviewed in the previous chapter) and in the analysis techniques adopted for extracting information from these data useful for operational applications. The modern technology resulted in the definition of different kinds of sensors for Earth observation based on different principles and with different properties. In this context, the challenging properties of new generation of sensors require the definition of novel data analysis methods. In this chapter we focus our attention on RS image classification methodologies, which are devoted to translate the features that represent the information present in the data in thematic maps representing land cover types according to the solution of a pattern recognition problem. In particular, we concentrate our attention on supervised classification algorithms, which require the availability of labeled samples

This chapter was published as chapter 3.2 of the book "Handbook of Pattern Recognition and Computer Vision", vol. 4, edited by Prof. C.H. Chen. Authors: L. Bruzzone, C. Persello.

for the training of the classification model. In this context, the availability of last generation RS images allowed the development of new applications that require the mapping of the Earth surface with high geometric precision and a high level of thematic details. However, the huge amount of data associated with these images requires the development of sophisticated automatic classification techniques capable to obtain accurate land-cover maps in a reasonable processing time.

In the last decades, a great effort has been devoted to exploit machine learning methods for classification of RS images. This has been done by introducing the use of neural networks (NN) in RS (with the pioneering work presented in [1]) for solving many different classification tasks. Several different paradigms and models of NN have been used in recent years for addressing remote sensing image classification problems, ranging from standard Multilayer Perceptron (MLP) network [1]-[3], to Radial Basis Functions (RBF) neural network [4], [5], structured neural networks [6] and hybrid architectures. Also more complex and structured architecture have been exploited for solving specific problems, like compound classification of multitemporal data [7], multiple classification systems made up of neural algorithms [8], [9], etc. All these methods share as common property the idea to perform the learning of the classification algorithm according to the minimization of the empirical risk, associated to the errors on the training set. However, the last frontiers of machine learning classifiers in RS are represented by methods based on the structural risk minimization principle (which allows one to effectively tune the tradeoff between empirical risk and generalization capability) rather than on the empirical risk minimization. The related statistical learning theory (formulated from Vapnik [10]) is at the basis of the support vector machine (SVM) classification approach. SVM is a classification technique based on kernel methods that has been proved very effective in solving complex classification problems in many different application domains. In the last few years, SVM gained a significant credit also in RS applications. The pioneering work of Gualtieri in 1998 [11] related to the use of SVM for classification of hyperspectral images has been followed from several different experiences of other researchers that analyzed the theoretical properties and the empirical performances of SVM applied to different kinds of classification problems [12]-[28]. The investigations include classification of hyperspectral data [11]-[18], multispectral images [19]-[26], VHR images [27], as well as multisource and multisensor classification scenarios [28]-[30]. SVMs revealed to be very effective classifiers and currently they are among the most adequate techniques for the analysis of last generation of RS data.

In all these cases the success of SVMs is due to the important properties of this approach, which integrated with the effectiveness of the classification procedure and the elegance of the theoretical developments, result in a very solid classification methodology in many different RS data classification domains. As it will be explained in the following section, this mainly depends on the fact that SVMs implement a classification strategy that exploits a margin-based “geometrical” criterion rather than a purely “statistical” criterion. In other words, SVMs do not require an estimation of the statistical distributions of classes to carry out the classification task, but they define the classification model by exploiting the concept of margin maximization.

The main properties that make SVM particularly attractive in RS applications can be summarized as follows [31]-[33]:

- their intrinsic effectiveness with respect to traditional classifiers thanks to the structural risk minimization principle, which results in high classification accuracies and very good generalization capabilities (especially in classification problems defined in high dimensional feature spaces and with few training samples, which is a typical situation in the classification of last generation of RS images);

- the possibility to exploit the kernel trick to solve non-linear separable classification problems by projecting the data into a high dimensional feature space and separating the data with a simple linear function;
- the convexity of the objective function used in the learning of the classifier, which results in the possibility to solve the learning process according to linearly constrained quadratic programming (QP) characterized from a unique solution (i.e., the system cannot fall into suboptimal solutions associated with local minima);
- the possibility of representing the convex optimization problem in a dual formulation, where only non-zero Lagrange multipliers are necessary for defining the separation hyperplane (which is a very important advantage in the case of large data sets). This is related to the property of sparseness of the solution.

Moreover, SVMs exhibit important advantages with respect to NN approaches. Among the others we recall: 1) higher generalization capability and robustness to the Hughes phenomenon; 2) lower effort required for the model selection in the learning phase (i.e., they involve less control parameters and thus computational time for their optimum values selection) and the implicit automatic architecture definition; 3) optimality of the solution obtained by the learning algorithm.

The objective of this chapter is to review the state of the art of SVM for the classification of RS data. In particular, Section 2.2 recalls the basic principles of SVM for pattern classification. Section 2.3 presents a literature survey about the most relevant papers that report studies about the application of SVM to the classification of different kinds of RS images and papers that propose advanced systems based on the SVM approach for the analysis of RS data. Along with this state-of-the-art review, we discuss about the operative adoption of SVM for the analysis of RS images and the direction of the future research on this topic. Finally, section 2.4 draws the conclusion of the chapter.

2.2 Support vector machine classifiers

Let us consider the problem of supervised classification of a generic *d*-dimensional image $\mathcal{I}$ of size $I \times J$ pixel. Let us assume that a training set $T = \{\mathcal{X}, \mathcal{Y}\}$ made up of *N* pairs $\left(\mathbf{x}_i, y_i\right)_{i=1}^{N}$ is available, where $\mathcal{X} = \{\mathbf{x}_i \mid \mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^{N} \subset \mathcal{I}$ is a subset of $\mathcal{I}$ and $\mathcal{Y} = \{y_i\}_{i=1}^{N}$ is the corresponding set of labels. For the sake of simplicity, since SVMs are binary classifiers, we first focus the attention on the two-class case (the general multiclass case will be addressed later). Accordingly, let us assume that $y_i \in \{+1; -1\}$ is the binary label of the pattern $\mathbf{x}_i$. The goal of the binary SVM is to divide the *d*-dimensional feature space in two subspaces, one for each class, through a separating hyperplane $H : y = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b = 0$. The final decision rule used to find the membership of a test sample is based on the sign of the discrimination function $f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b$ associated to the hyperplane. Therefore, a generic pattern $\mathbf{x}$ will be labeled according to the following rule:

$$\begin{aligned} f(\mathbf{x}) > 0 \quad &\Rightarrow \quad \mathbf{x} \in \mathit{class}\ +1 \\ f(\mathbf{x}) \leq 0 \quad &\Rightarrow \quad \mathbf{x} \in \mathit{class}\ -1 \end{aligned} \tag{2.1}$$

The training of an SVM consists in finding the position of the hyperplane *H*, estimating the values of the vector $\mathbf{w}$ and the scalar *b*, according to the solution of an optimization problem. From a geometrical point of view, $\mathbf{w}$ is a vector perpendicular to the hyperplane *H* and thus defines its orientation. The distance of the *H* to the origin is $b/\|\mathbf{w}\|$, while the distance of a sample $\mathbf{x}$ to the hyperplane is $f(\mathbf{x})/\|\mathbf{w}\|$. Let us define the *functional margin* $F=\min\{y_i f(\mathbf{x}_i)\}$, $i=1,...,N$ and the *geometric margin* $g=F/\|\mathbf{w}\|$. The geometric margin represents the minimum Euclidean distance between the available training samples and the hyperplane.

2.2.1 Training of linear SVM - maximal margin algorithm.

In the case of a linearly separable problems, the learning of an SVM can be performed with the maximal margin algorithm, which consists in finding the hyperplane *H* that maximizes the geometric margin *G*. Rescaling the hyperplane parameters $\mathbf{w}$ and *b* such that the functional margin $F=1$, it turns out that the optimal hyperplane can be determined as the solution of the following convex quadratic programming problem:

$$\begin{cases} \min\limits_{\mathbf{w},b}: \dfrac{1}{2}\|\mathbf{w}\|^2 \\ y_i \cdot \left[\langle \mathbf{w}\cdot\mathbf{x}_i\rangle + b\right] \geq 1, \quad \forall i=1,\ldots,N \end{cases} \tag{2.2}$$

Let H_1 and H_2 be two hyperplane parallel to the separating hyperplane *H* and equidistant from it:

$$\begin{aligned} H_1 &: f(\mathbf{x}) = \langle \mathbf{w}\cdot\mathbf{x}\rangle + b = +1 \\ H_2 &: f(\mathbf{x}) = \langle \mathbf{w}\cdot\mathbf{x}\rangle + b = -1 \end{aligned} \tag{2.3}$$

The goal of the training phase is to find the values of $\mathbf{w}$ and *b* such that the geometric distance between H_1 and H_2 is maximized with the condition that there is no sample between them. Since direct handling of inequality constraints is difficult, Lagrange theory is usually exploited by introducing Lagrange multipliers $\alpha_{i=1}^{N}$ for the quadratic optimization problem. This leads to an alternative dual representation:

$$\begin{cases} \max\limits_{\alpha}: \left\{ \sum_{i=1}^{N}\alpha_i - \dfrac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j\rangle \right\} \\ \sum_{i=1}^{N} y_i\alpha_i = 0, \;\; \alpha_i \geq 0, \quad 1\leq i \leq N \end{cases} \tag{2.4}$$

The Karush–Kuhn–Tucker (KKT) complementarity conditions provide useful information about the structure of the solution. They state that the optimal solution $\boldsymbol{\alpha}^*$, $(\mathbf{w}^*, b^*)$ should satisfy:

$$\alpha_i^*[y_i(\langle \mathbf{w}^*\cdot\mathbf{x}_i\rangle + b^*) - 1] = 0, \qquad i=1,....,N \tag{2.5}$$

This implies that only input samples $\mathbf{x}_i$ for which the functional margin is one (and that therefore lie closest to the hyperplane, i.e., lie on H_1 or H_2) are associated to Lagrange multipliers $\alpha_i > 0$. All the other multipliers α_i^* are zero. Hence, only these samples are involved in the expression for the weight vector. It is for this reason that they are called *support vectors* (SV). Thus we can write that $\mathbf{w}^* = \sum_{i=1}^{N} y_i \alpha_i^* \mathbf{x}_i = \sum_{i\in SV} y_i \alpha_i^* \mathbf{x}_i$. It is worth noting that the term *b* does not appear in the dual problem, and should be calculated making use of the primal constraints:

$$b = -\frac{\max_{y_i=-1}(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle) + \min_{y_i=+1}(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle)}{2} \tag{2.6}$$

Once the values for $\mathbf{w}$ and b are determined by solving the optimization problem, one generic test sample is classified on the basis of the sign of the discriminant function, that can be expressed as:

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b = \left(\sum_{i \in SV} y_i \alpha_i^* \mathbf{x}_i \right) \cdot \mathbf{x} + b = \sum_{i \in SV} y_i \alpha_i^* \langle \mathbf{x}_i \cdot \mathbf{x} \rangle + b\,. \tag{2.7}$$

Note that the training samples appear only in the form of dot product. This property of the dual form will be exploited later to extend the formulation to nonlinear problems.

2.2.2 Training of linear SVM - soft margin algorithm.

The maximum margin training algorithm can not be used in many real world problems where the available training samples are not linearly separable because of noisy samples and outliers (this is very common in real RS classification problems). In these cases, the soft margin algorithm is used in order to handle nonlinear separable data. This is done by defining the so called slack variables as:

$$\xi[(\mathbf{x}_i, y_i), (\mathbf{w}, b)] = \xi_i = \max[0, 1 - y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b)] \tag{2.8}$$

Slack variables allow one to control the penalty associated with misclassified samples. In this way the learning algorithm is robust to both noise and outliers present in the training set, thus resulting in high generalization capability. The optimization problem can be formulated as follows:

$$\begin{cases} \min\limits_{\mathbf{w},b} \left\{ \dfrac{1}{2} \|\mathbf{w}\|^2 + C \sum\limits_{i=1}^{N} \xi_i \right\} \\ y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ \forall i = 1, \ldots, N \end{cases} \tag{2.9}$$

where $C \geq 0$ is the regularization parameter that allows one to control the penalty associated to errors (if $C = \infty$ we come back to the maximal margin algorithm), and thus to control the trade-off between the number of allowed mislabeled training samples and the width of the margin. If the value of *C* is too small, many errors are permitted and the resulting discriminant function will poorly fit with the data; on the opposite, if *C* is too large, the classifier may overfit the data instances, thus resulting in low generalization ability. A precise definition of the value of the *C* parameter is crucial for the accuracy that can be obtained in the classification step and should be derived through an accurate model selection phase.

Similarly to the case of the maximal margin algorithm, the optimization problem (2.9) can be rewritten in an equivalent dual form:

$$\begin{cases} \max\limits_{\alpha} : \left\{ \sum\limits_{i=1}^{N} \alpha_i - \dfrac{1}{2} \sum\limits_{i=1}^{N} \sum\limits_{j=1}^{N} y_i y_j \alpha_i \alpha_j \left\langle \mathbf{x}_i, \mathbf{x}_j \right\rangle \right\} \\ \sum_{i=1}^{N} y_i \alpha_i = 0,\ 0 \leq \alpha_i \leq C,\ \ 1 \leq i \leq N \end{cases} \tag{2.10}$$

Note that the only difference between (2.10) and (2.4) is in the constraint on the multipliers $\{\alpha_i\}_{i=1}^{N}$ that for the soft margin algorithm are bounded by the parameter *C*. For this reason this problem is also known as box constrained problem. The KKT conditions become in this case:

$$\begin{cases} \alpha_i \left[y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) - 1 + \xi_i \right] = 0, & i = 1, ..., l \\ \xi_i(\alpha_i - C) = 0, & i = 1, ..., l \end{cases} \tag{2.11}$$

Varying the values of the multipliers $\{\alpha_i\}_{i=1}^{N}$ three cases can be distinguished:

1. if $\alpha_i = 0 \Rightarrow \xi_i = 0$ and $y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1$;
2. if $0 < \alpha_i < C$, we have that $y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) + \xi_i = 1$, but given that $\xi_i = 0$ we have that $y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) = 1$;
3. if α_i = C, $\Rightarrow y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) + \xi_i = 1$, but given that $\xi_i \geq 0$ we have that $y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \leq 1$.

The KKT conditions can therefore be rewritten as:

$$\begin{cases} \alpha_i = 0 & \Rightarrow y_i f(\mathbf{x}_i) \geq 1 \\ 0 < \alpha_i < C_i & \Rightarrow y_i f(\mathbf{x}_i) = 1 \\ \alpha_i = C_i & \Rightarrow y_i f(\mathbf{x}_i) \leq 1 \end{cases} \tag{2.12}$$

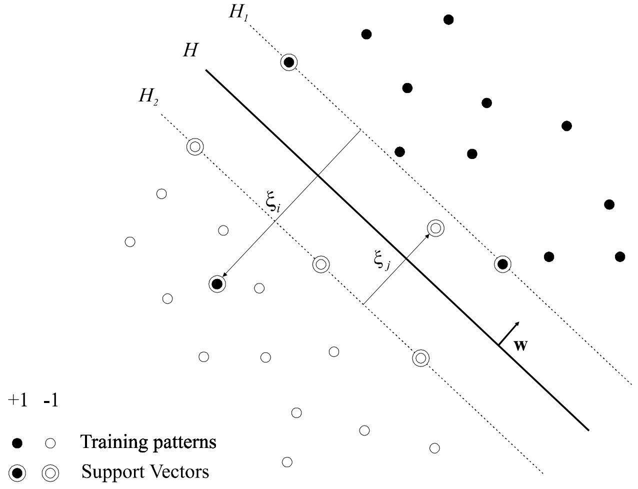


Fig. 2.1. Qualitative example of a separating hyperplane in the case of a non linear separable classification problem.

The *support vectors* with multiplier $\alpha_i = C$ are called *bound support vectors* (BSV) and are associated to slack variables $\xi_i \geq 0$; the ones with $0 < \alpha_i < C_i$ are called *non bound support vectors* (NBSV) and lie on the margin hyperplane H_1 or H_2 ($y_i f(\mathbf{x}_i) = 1$).

2.2.3 Training of non linear SVM - kernel trick.

An important improvement to the above-described methods consists in considering non linear discriminant functions for separating the two information classes. This can be obtained by transforming the input data into a high dimension (Hilbert) feature space $\Phi(\mathbf{x}) \in \mathbb{R}^{d'}$ ($d' > d$) where the transformed samples can be better separated by a hyperplane. The main problem is to explicitly choose and calculate the function $\Phi(\mathbf{x}) \in \mathbb{R}^{d'}$ for each training samples. But given that the input points in dual formulation [see (2.10)] appear in the form of inner products, we can do this mapping in an implicit way by exploiting the so called kernel trick. Kernel methods provide an elegant and effective way of dealing with this problem by replacing the inner product in the input space with a kernel function such that:

$$K(x_i, x_j) = \langle(\phi(\mathbf{x}_i)\cdot\phi(\mathbf{x}_j)\rangle \qquad i, j = 1, \dots, N \tag{2.13}$$

implicitly calculating the inner product in the transformed space.

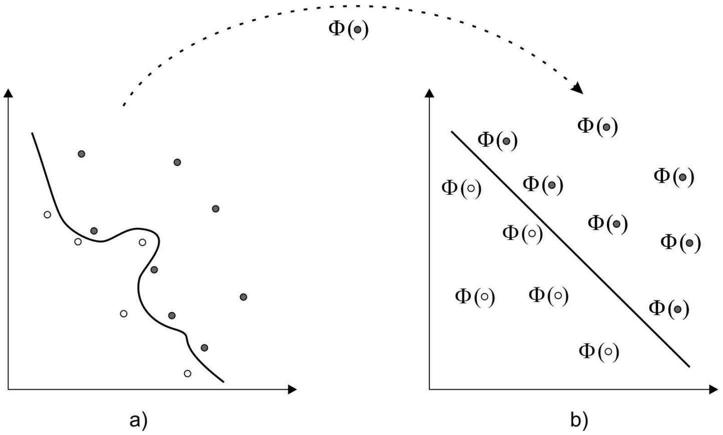


Fig. 2.2. Transformation of the input data by means of a kernel function into a high dimension feature space. a) Input feature space; b) kernel induced high dimensional feature space.

The soft margin algorithm for nonlinear function can be represented by the following optimization problem:

$$\begin{cases} \max\limits_{\alpha}\left\{\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j)\right\} \\ \sum_{i=1}^{N} y_i \alpha_i = 0, \ 0 \le \alpha_i \le C \text{ and } 1 \le i \le N \end{cases} \tag{2.14}$$

And the discrimination function becomes:

$$f(\mathbf{x}) = \sum_{i \in SV} y_i \alpha_i^* k(\mathbf{x}_i \cdot \mathbf{x}) + b \tag{2.15}$$

The condition for a function to be a valid kernel is given by the Mercer’s theorem [32]. The most widely used non-linear kernel functions are the followings [31]:

- homogeneous polynomial function: $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)^p, \quad p \in \mathbb{Z}$
- inhomogeneous polynomial function: $k(\mathbf{x}_i, \mathbf{x}_j) = (c + (\mathbf{x}_i \cdot \mathbf{x}_j))^p, \quad p \in \mathbb{Z}, c \ge 0$
- Gaussian function: $k(\mathbf{x}_i, \mathbf{x}_j) = e^{\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}, \qquad \sigma \in \mathbb{R}$

2.2.4 Multiclass architectures

As stated in the previous section, SVMs are binary classifiers. However, several strategies have been proposed to address multiclass problems with SVMs. Let $\Omega = \{\omega_1, \dots, \omega_L\}$ be the set of L information classes associated with the different land cover types present in the study area. In order to define a multiclass architecture based on different binary classifiers, the general approach consists of: 1) defining an ensemble of binary classifiers; and 2) combining them according to some decision rules. The definition of the ensemble of binary classifiers involves the definition of a set of two-class problems, each modeled with two groups Ω_A and Ω_B of classes. The selection of these subsets depends on the kind of approach adopted to combine the ensemble. In the following, we describe the two most widely adopted (parallel) multiclass strategies, i.e., the *One-Agains-All* (OAA) and *One-Against-One* (OAO) strategies.

1) *One-Against-All:* the one-against-all (OAA) strategy represents the earliest and one of the most common multiclass approach used for SVMs. It involves a parallel architecture made up of *L* SVMs, one for each class (Fig. 2.3). Each SVM solves a two-class problem defined by one information class against all the others, i.e.,

$$\begin{cases} \Omega_A = \omega_i \\ \Omega_B = \Omega - \omega_i \end{cases} \tag{2.16}$$

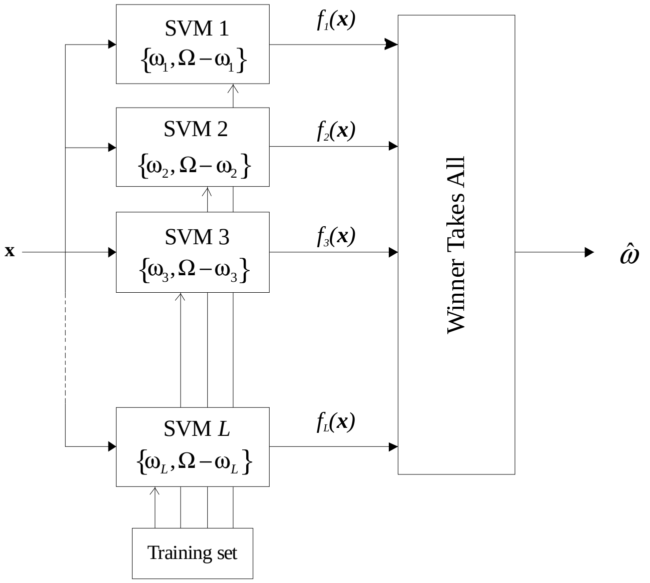

SVM 1
$\{\omega_1, \Omega - \omega_1\}$
SVM 2
$\{\omega_2, \Omega - \omega_2\}$
SVM 3
$\{\omega_3, \Omega - \omega_3\}$
SVM L
$\{\omega_L, \Omega - \omega_L\}$
$f_1(x)$
$f_2(x)$
$f_3(x)$
$f_L(x)$
x
Winner Takes All
$\hat{\omega}$
Training set

Fig. 2.3 Block diagram of the *One-Against-All* multiclass architecture

The *winner-takes-all* rule is used for the final decision, i.e., the winning class is the one corresponding to the SVM with the highest output (discriminant function value).

2) *One-Against-One:* the main problem of the OAA strategy is that the discrimination between an information class and all the others often leads to the estimation of complex discriminant functions. In addition, a problem with strongly unbalanced prior probabilities should be solved by each SVM. The idea behind the *one-against-one* (OAO) strategy is that of a different reasoning, in which simple classification tasks are made possible thanks to a parallel architecture made up of a large number of SVMs. The OAO strategy involves $L(L-1)/2$ SVMs, which model all possible pairwise classifications. In this case, each SVM carries out a binary classification in which two information classes ω_i and ω_j are analyzed against each other by means of a discriminant function $f_{ij}(\mathbf{x})$. Consequently, the grouping becomes:

$$\begin{cases} \Omega_A = \omega_i \\ \Omega_B = \omega_j \end{cases} \tag{2.17}$$

Before the decision process, it is necessary to compute for each class $\omega_i \in \Omega$ a score function $D_i(\mathbf{x})$, which sums the favorable and unfavorable votes expressed for the considered class

$$D_i(\mathbf{x}) = \sum_{\substack{j=1 \\ j\neq i}}^{L} \operatorname{sgn}[f_{ij}(\mathbf{x})] \tag{2.18}$$

The final decision in the OAO strategy is taken on the basis of the *winner-takes-all* rule, which corresponds to the following maximization:

$$\mathbf{x} \in \omega \Leftrightarrow \omega = \arg\max_{i=1,\dots,L} \{D_i(\mathbf{x})\} \tag{2.19}$$

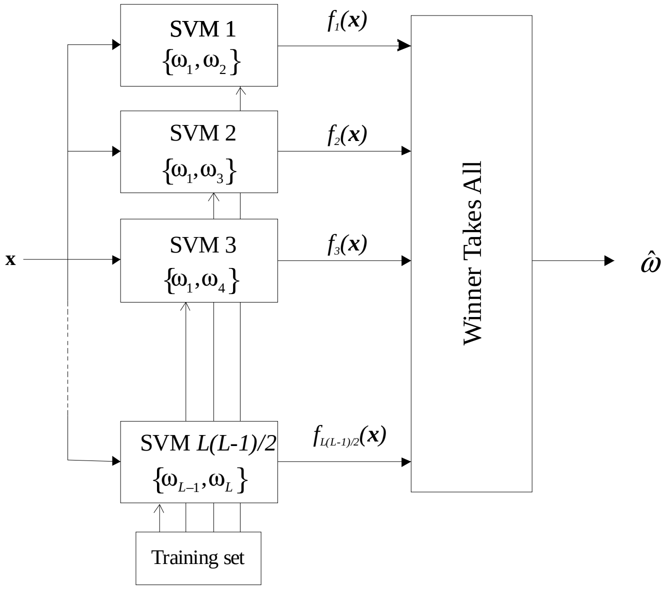


Fig. 2.4 Block diagram of the *One-Against-One* multiclass architecture

Other multiclass architectures proposed in the literature are the Directed Acyclic Graph SVM (DAGSVM) [34] and different approaches based on binary hierarchical trees (BHT) [16], [35].

2.3 SVM for the classification of RS data

In the last decade many studies have been published in the RS literature on the application of SVM classifiers to the analysis of RS data. Table 2.1 (which is not exhaustive) presents some relevant papers about the applications of SVM to the classification of RS data, providing a short description of the study and the kind of data used for the experimental analysis. The SVM approach has been first applied to the classification of hyperspectral data [11], which require the classifier to operate in large dimensional feature spaces. Supervised classification of hyperspectral images is a very complex methodological problem due to many different issues, among which we recall the typical small value of the ratio between the number of training samples and the number of available spectral channels, which results in the so-called course of dimensionality

(Hughes phenomenon) [36]. Thanks to the structural risk minimization principle and the margin-based approach, SVMs represent an effective choice for the classification of this specific kind of data. Several papers [11]-[18] confirm the effectiveness of SVMs in the classification of hyperspectral images, which outperform other classification algorithms both in terms of classification accuracy and generalization ability. In particular, in [16] it is found that SVMs are much more effective than other conventional nonparametric classifiers (i.e., the RBF neural networks and the *k*-NN classifier) in terms of classification accuracy, computational time, stability to parameter setting, and generalization ability. In [15], the SVM approach was compared with neural networks and fuzzy methods on six hyperspectral images acquired with the 128-band HyMap spectrometer. The authors of the study concluded that SVMs yield better outcomes than neural networks regarding accuracy, simplicity, and robustness. In [17], SVMs were compared with other kernel-based methods, i.e., with regularized radial basis function NN, kernel Fisher discriminant analysis, and regularized AdaBoost. The results obtained on an AVIRIS data set show that SVMs are more beneficial, yielding better results than other kernel-based methods, ensuring sparsity and lower computational cost.

Nevertheless, SVMs revealed adequate for the analysis of many different kinds of RS data, i.e., multispectral imagery and SAR imagery (with different resolutions) and LIDAR data. Several papers present a comparison between SVM and other supervised algorithms applied to the classification of different kinds of RS images [20], [23], [25], [30]. In [20], for instance, the authors compared the accuracies obtained by the classification of a Landsat Thematic mapper (TM) scene with four different supervised classifiers, i.e., SVM, maximum likelihood (ML), MLP neural networks (NN), and decision tree classifier (DTC). The obtained results show that SVM was in general sharply more accurate than ML and DTC, and more accurate than NN in most of the cases. In [21], the SVM algorithm was applied to the classification of ASTER data acquired in an urban area of Beer Sheva, Israel. Field validations show that the classification is reliable for urban studies with high classification accuracy. In [23], the SVM classifier, as well as the well-known ML classifier and a context-based classifier based on Markov random fields, were applied to the automatic land cover classification of a Landsat TM image taken on the Tenerife Island. The authors found that SVM was more accurate than the other classification algorithms, but the classification map was not completely satisfying when investigated visually. In the experimental analysis conducted in [25], it is observed that SVM leaded to slightly higher classification accuracies than (MLP) NN. For both classifiers, the accuracy depends on factors such as the number of hidden nodes in the case of NN, and kernel parameters in the case of SVM. Thus, the model selection phase is fundamental for obtaining good results, but the training time required by the SVM is less than the one taken by NN.

SVM can be particularly effective also in the analysis of very high resolution (VHR) images. The typical poor spectral resolution of VHR images requires the extraction of additional features (e.g., texture and geometric measures) to characterize the objects present in the scene under investigation and to discriminate different land-cover classes. Different features modeling objects at different scales are generally necessary for an adequate characterization of the information classes [27], thus resulting in classification problems characterized by large dimensional feature spaces (with some analogies with the problems related to the classification of hyperspectral image). The study proposed in [27] points out that SVM can be effectively applied to the classification of VHR images using a feature extraction block that aims at adaptively modeling the spatial context of each pixel according to a hierarchical multilevel segmentation of the scene. A similar approach can also be adopted for the joint classification of SAR and optical data with SVM, as presented in [29]. In [30], an analysis is proposed on the joint use of hyperspectral and LIDAR data for the classification of complex forest areas. The experimental results obtained in [29]-[30] show that SVMs are effective for combining multisensor data in complex classification problems and outperforms other more traditional classifiers.

Table 2.1– Selected papers related to the application of SVM to the classification of different kinds of RS data

Authors	**Description**	**RS data**
J. A. Gualtieri and S. Chettri [13]	In this paper, the authors introduce SVM for the classification of RS data. In particular they applied SVM to hyperspectral data acquired by NASA's AVIRIS sensor and the commercially available AISA sensor. The authors discuss the robustness of SVM to the course of dimensionality (Hughes phenomenon).	AVIRIS (224 spectral bands) and AISA (20-40 bands)
F. Melgani, L. Bruzzone [16]	This paper addresses the problem of the classification of hyperspectral RS images by SVMs. The authors propose a theoretical discussion and experimental analysis aimed at understanding and assessing the potentialities of SVM classifiers in hyperdimensional feature spaces. Then, they assess the effectiveness of SVMs with respect to conventional feature-reduction-based approaches and their performances in hypersubspaces of various dimensionalities. To sustain such an analysis, the performances of SVMs are compared with those of two other nonparametric classifiers (i.e., radial basis function neural networks and the K-nearest neighbor classifier). Four different multiclass strategies are analyzed and compared: the one-against-all, the one-against-one, and two hierarchical tree-based strategies.	AVIRIS (224 spectral bands)
G. Camps-Valls, L. Bruzzone [17]	This paper presents the framework of kernel-based methods in the context of hyperspectral image classification, illustrating from a general viewpoint the main characteristics of different kernel-based approaches and analyzing their properties in the hyperspectral domain. In particular, the performances of the following techniques are assessed: regularized radial basis function neural networks (Reg-RBFNN), standard support vector machines (SVMs), kernel Fisher discriminant (KFD) analysis, and regularized AdaBoost (Reg-AB).	AVIRIS (224 spectral bands)

Authors	**Description**	**RS data**
G..M. Foody, A. Mathur [19]	In this paper, an approach for multiclass classification of airborne sensor data by a single SVM analysis is evaluated against a series of classifiers that are widely used in RS, with particular regard to the effect of training set size on classification accuracy. In addition to the SVM, the same data sets are classified using discriminant analysis, decision tree, and multilayer perceptron neural network. For each classification technique, the accuracy is positively related with the size of the training set. In general, the most accurate classifications are obtained with the SVM approach.	Airborne Thematic Mapper (ATM) (11 spectral bands, spatial resolution of 5m)
C. Huang, L.S. Davis, J.R.G. Townshend [20]	This paper introduces the theory of SVM and provides an experimental evaluation of its accuracy, stability, and training speed in deriving land cover classifications from satellite images. SVM algorithm is compared with other supervised algorithms: maximimum likelihood (ML) classifier, neural network classifier, and decision tree classifier.	(Spatially degraded) Landsat Thematic Mapper (TM)
G. Zhu, D. G. Blumberg [21]	This paper presents a study on the mapping of urban environments using ASTER data and SVM-based classification algorithms. A case study of the classification of the area of Beer Sheva, Israel is presented. Field validation shows that the classification is reliable and precise.	Advanced Spaceborne Thermal Emission and Reflectance Radiometer (ASTER)
L. Su, M. J. Chopping, A. Rango, J. V. Martonchik, D. P. C. Peters [22]	This paper present a study on mapping and monitoring the desert environment using SVM for the analysis of Multi-angle Imaging Spectro-Radiometer (MISR) RS data. Many classification experiments are performed to find the optimal combination of MISR multi-angle data for maximizing the classification accuracy.	Multi-angle Imaging Spectro-Radiometer (MISR)
J. Keuchel, S. Naumann, M. Heiler, A. Siegmund [23]	This paper presents three different approaches to the classification of satellites images: maximum likelihood classifier, SVM, and iterated conditional model (ICM) to perform contextual classification using Markov random field model. The classification algorithms are applied to a Landsat 5 TM image of Tenerife, the largest of the canary Island.	Landsat 5 TM
B. Dixon, N. Candade [25]	This paper presents a study on the comparison between SVM and NN for the classification of RS data. An experimental analysis is carried on Landsat 5 TM data, acquired in the South West of Florida. The obtained results confirm that SVM and NN outperform the traditional ML classifier. SVM classification results slightly more accurate than NN requiring much less computational effort in the training phase.	Landsat 5 TM

Authors	Description	RS data
L. Bruzzone, L. Carlin [27]	This paper proposes a system for the classification of VHR images. The proposed system is made up of two main blocks: 1) a feature-extraction block that aims at adaptively model the spatial context of each pixel according to a hierarchical multilevel segmentation of the scene and 2) a classification block based on SVM. Experimental results obtained on VHR images confirm the effectiveness of the proposed system.	Quickbird
B. Waske, S. Van der Linden [29]	This paper presents a strategy for the joint classification of multiple segmentation levels from multisensor imagery, using SAR and optical data. The two data sets are separately segmented at different scale levels and independently classified by two SVM-based classifiers. The fusion strategy is based on the application of an additional classifier, which takes in input the soft output of the pre-classified results of the two data sets. The obtained experimental results show that the useful combination of multilevel-multisensor data is feasible with machine learning techniques like SVM and Random forest.	Multitemporal SAR data and Landsat 5 TM
M. Dalponte, L. Bruzzone, and D. Gianelle [30]	In this paper, the authors propose an analysis on the joint use of hyperspectral and light detection and ranging (LIDAR) data for the classification of complex forest areas. In greater detail, they present: 1) an advanced system for the joint use of hyperspectral and LIDAR data in complex classification problems; 2) an investigation on the effectiveness of the very promising SVM and Gaussian ML with leave-one-out covariance algorithm for the analysis of forest areas characterized from a high number of species; and 3) an analysis of the effectiveness of different LIDAR returns and channels for increasing the classification accuracy obtained with hyperspectral images.	Hyperspectral (126 spectral bands) and LIDAR (mean density of 5.6 points per square meter)

The RS literature related to SVM is not limited to the use of this approach on different data and different application domains. Recently, more advanced SVM-based classifiers have been developed for facing complex problems related to the properties of RS images. A list of relevant papers that introduced advanced techniques based on SVM for the classification of RS data is reported Table 2.2. These papers represent the most recent (and in some cases on-going) research activities in this field and give insight about the research direction for the next years.

In this context, it is worth mentioning the semi-supervised SVM classifiers [37]-[43], which are devised for addressing ill-posed problems characterized by a very small ratio between the number of available training samples and the number of features by reinforcing the learning procedure with the use of unlabeled samples. It is worth noting, that even if SVMs have very good generalization capability, they cannot model the classification problem when very few training samples are available ("strongly" ill-posed problems). In these cases, the exploitation of the unlabeled samples to enrich the information of the training samples can result in a significant improvement in the model estimation. The first work on semisupervised SVM in RS was presented in [37], [38]. The presented semisupervised SVM (S^{3}VM) is based on transductive inference that exploits a specific iterative algorithm which gradually searches a reliable separating hyperplane in the kernel space with a process that incorporates both labeled and unlabeled samples in the training phase. In [39], an S^{3}VM classification technique is proposed, where the learning phase is performed by optimizing the objective function directly in the primal formulation (without exploiting the dual representation that can be obtained with Lagrange multipliers). In [40], the Laplacian SVM technique [41] is introduced in the RS community. This technique adopts an additional regularization term on the geometry of both labeled and unlabeled samples by using the graph Laplacian. This method follows a non-iterative optimization procedure in contrast to most transductive learning methods and provides out-of-sample predictions in contrast to graph-based approaches. Experimental results confirm the effectiveness of S^{3}VM techniques for solving ill-posed RS classification problems. In general S^{3}VM provides higher accuracy and better generalization ability than standard supervised SVM. In this respect, a more detailed picture of the status on the research on the application of S^{3}VM to hypedimensional problems can be found in [43].

Other studies address the inclusion of the spatial-context information of the single pixel in the SVM classification process. To this end, [44] proposes a framework for applying the maximum a posteriori (MAP) estimation principle in remote sensing image segmentation, which incorporates contextual and geometrical information in the SVM classification process by means of Markov random field (MRF). In [45], the use of composite kernels is introduced in remote sensing to adopt different kernel functions for different subsets of features to combine spatial and spectral information in an effective way. In [47], a context-sensitive semisupervised SVM is proposed, which exploits the contextual information of the pixels during the learning phase, in order to improve the robustness to possible mislabeled training patterns (which are not unlikely to be present in the reference data due to different kinds of errors that may occur in the collection of labeled samples). For details we refer the reader to chapter 4 of this dissertation.

The study in [48] addresses the problem of automatic updating the land-cover maps by using RS images periodically acquired over the same investigated area under the hypothesis that a reliable ground truth is not available for all the considered acquisitions. The problem is modeled under the domain-adaptation framework by introducing a novel method designed for land-cover map updating, which is based on a domain-adaptation SVM (DASVM) technique. Given two RS images I_1 and I_2 acquired over the same area at different times (t_1 and t_2, respectively), the goal of the DASVM is to obtain an accurate classification of I_2 by exploiting the labeled training samples from reference image I_1 and the unlabeled samples from the new image I_2. The DASVM algorithm is based on an iterative process, which starts by training an SVM classifier with the original training samples of I_1 and gradually introduces semilabeled samples of I_2 and erases the original training samples. At convergence a final classification function ruled only by semilabeled samples at time t_2 is obtained. In addition, the authors propose a circular accuracy assessment strategy for the validation of the results obtained by domain-adaptation classifiers when no reference data for the considered image I_2 are available.

Another recent and promising approach to the analysis RS data is associated with active learning [49]-[50], which allows an interactive classification of RS images (see chapter 5). The active learning approach is based on the iteration on three different conceptual steps. In the first step the learning process queries unlabeled samples to select the most informative ones; in the second step the supervisor (e.g., the user) provides a label to the selected samples interacting with the system; and in the third step the learner updates the classification rule by retraining with the updated training set. In [49], it is noted that SVMs are particularly suited to active learning since they are characterized by a small set of support vectors (SVs) which can be easily updated over successive learning iterations. Moreover, one of the most efficient query functions is based on the selection of the sample closest to the separating hyperplane defined at the considered iteration. For additional information about recent developments in kernel methods for the analysis of RS images, we refer the reader to [51]. For more details on this topic we refer the reader to chapter 5 of this thesis.

Table 2.2 – Relevant papers about advanced techniques based on SVM for the classification of RS data.

Authors	**Description**
L. Bruzzone, M. Chi, M. Marconcini [38]	This paper introduces a semisupervised classification method that exploits both labeled and unlabeled samples for addressing ill-posed problems with SVMs. The proposed method exploit specific iterative algorithms which gradually search a reliable separating hyperplane in the kernel space with a process that incorporates both labeled and unlabeled samples in the training phase. The authors propose a novel modified transductive SVM classifier designed for addressing ill-posed RS problems, which has the following properties: 1) it is based on a novel transductive procedure that exploits a weighting strategy for unlabeled patterns, based on a time-dependent criterion; 2) is able to mitigate the effects of suboptimal model selection (which is unavoidable in the presence of small-size training sets); and 3) can address multiclass cases.
M. Chi, L. Bruzzone [39]	This paper addresses classification of hyperspectral RS images with kernel-based methods defined in the framework of semisupervised SVM (S^3VMs). In particular, the authors analyzed the critical problem of the nonconvexity of the cost function associated with the learning phase of S^3VMs by considering different (S^3VMs) techniques that solve optimization directly in the primal formulation of the objective function. As the nonconvex cost function can be characterized by many local minima, different optimization techniques may lead to different classification results. The presented techniques are compared with S^3VMs implemented in the dual formulation in the context of classification of real hyperspectral remote sensing images.
L. Gomez-Chova, G. Camps-Valls, J. Munoz-Mari, J. Calpe [40]	This letter presents a semisupervised method based on kernel machines and graph theory for RS image classification. The SVM is regularized with the unnormalized graph Laplacian, thus leading to the Laplacian SVM (LapSVM). The method is tested in the challenging problems of urban monitoring and cloud screening, in which an adequate exploitation of the wealth of unlabeled samples is critical.

Authors	**Description**
A. A. Farag, R. M. Mohamed, A. El-Baz [44]	This paper proposes a complete framework for applying the maximum a posteriori (MAP) estimation principle in RS image segmentation. The MAP principle provides an estimate for the segmented image by maximizing the posterior probabilities of the classes defined in the image. The posterior probability can be represented as the product of the class conditional probability (CCP) and the class prior probability (CPP). For the CCP, a supervised algorithm which uses the SVM density estimation approach is proposed. For the CPP estimation, Markov random field (MRF) is a common choice which incorporates contextual and geometrical information in the estimation process.
G. Camp-Valls, L. Gomez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J Calpe-Maravilla [45]	This letter presents a framework of composite kernel machines for enhanced classification of hyperspectral images. This novel method exploits the properties of Mercer's kernels to construct a family of composite kernels that easily combine spatial and spectral information. This framework of composite kernels demonstrates: 1) enhanced classification accuracy as compared to traditional approaches that take into account the spectral information only: 2) flexibility to balance between the spatial and spectral information in the classifier; and 3) computational efficiency.
M. Marconcini, G. Camps-Valls, L. Bruzzone [46]	This letter presents a novel composite semisupervised SVM for the spectral–spatial classification of hyperspectral images. In particular, the proposed technique exploits the following: 1) unlabeled data for increasing the reliability of the training phase when few training samples are available and 2) composite kernel functions for simultaneously taking into account spectral and spatial information included in the considered image. Experiments carried out on a hyperspectral image pointed out the effectiveness of the presented technique, which resulted in a significant increase of the classification accuracy with respect to both supervised SVMs and progressive semisupervised SVMs with single kernels, as well as supervised SVMs with composite kernels.

Authors	Description
L. Bruzzone, C. Persello [47] (see chapter 4)	This paper presents a novel context-sensitive semisupervised SVM (CS^{4}VM) classifier, which is aimed at addressing classification problems where the available training set is not fully reliable, i.e., some labeled samples may be associated to the wrong information class (mislabeled patterns). Unlike standard context-sensitive methods, the proposed CS^{4}VM classifier exploits the contextual information of the pixels belonging to the neighborhood system of each training sample in the learning phase to improve the robustness to possible mislabeled training patterns. This is achieved according to both the design of a semisupervised procedure and the definition of a novel contextual term in the cost function associated with the learning of the classifier. In order to assess the effectiveness of the proposed CS^{4}VM and to understand the impact of the addressed problem in real applications, the authors also present an extensive experimental analysis carried out on training sets that include different percentages of mislabeled patterns having different distributions on the classes. In the analysis they also study the robustness to mislabeled training patterns of some widely used supervised and semisupervised classification algorithms (i.e., conventional SVM, progressive semisupervised SVM, Maximum Likelihood, and *k*-Nearest Neighbor)
L. Bruzzone, M. Marconcini [48]	In this paper, the authors address automatic updating of land-cover maps by using RS images periodically acquired over the same investigated area under the hypothesis that a reliable ground truth is not available for all the considered acquisitions. The problem is modeled in the domain-adaptation framework by introducing a novel method designed for land-cover map updating, which is based on a domain-adaptation SVM technique. In addition, a novel circular accuracy assessment strategy is proposed for the validation of the results obtained by domain-adaptation classifiers when no ground-truth labels for the considered image are available.
D. Tuia, F. Ratle, F. Pacifici, A. Pozdnoukhov, M. Kanevski, F. Del Frate, D. Solimini, W. J. Emery [50]	In this paper, an active learning method is proposed for the semi-automatic selection of training sets in RS image classification. The method adds iteratively to the current training set the unlabeled pixels for which the prediction of an ensemble of classifiers based on bagged training sets show maximum entropy. This way, the algorithm selects the pixels that are the most uncertain and that will improve the model if added in the training set. The user is asked to label such pixels at each iteration.

2.4 Discussion and conclusion

In this chapter, we presented a review on SVMs in the classification of RS data, recalling their theoretical formulation, and discussing the motivations at the basis of their use in RS. We presented a literature survey about the adoption of SVMs for the analysis of different kinds of RS images. We observed a large variety of studies published on the use of SVMs for the analysis of different kinds of RS data, which confirm that SVMs represent a valuable and effective tool for the analysis of RS data and can be used in many different applications in the context of RS.

We observed that one of the most appealing properties of SVM for the classification of RS data is its high generalization capability and robustness to the Hughes effect, which allow SVMs to operate in large dimensional feature spaces with few training samples. For this reason, SVMs represent an effective choice for the classification of hyperspectral data. Nevertheless, the SVM approach turned out to be particularly effective also in the classification of very high resolution (VHR) images, which typically require the extraction of several additional features to characterize and discriminate the different land-cover classes. Thus, both the classification of VHR and hyperspectral images typically result in classification problems characterized by large dimensional feature spaces. Moreover, thanks to its distribution-free approach and the capability to cope with strongly non-linear problems by means of the kernel function, SVMs are a valuable tool also for the classification of data acquired by different information sources.

In addition, we pointed out the most recent works about the development of advanced SVM-based techniques for the analysis of RS data. Among these developments, we recall semisupervised and domain-adaptation SVM, techniques based on SVM that exploit the spatial-context information, and active learning methods. Semisupervised SVMs have shown to be effective in exploiting both labeled and unlabeled samples for the learning of the classification algorithm, further augmenting the generalization capability and the robustness to the Hughes phenomenon with respect to standard supervised SVM. Domain-adaptation SVM resulted effective for addressing the problem of automatic updating land-cover maps by using RS images periodically acquired over the same investigated area. Context-sensitive techniques based on SVM have been proposed for both regularizing the classification map (exploiting the context information in the classification phase) or for improving the robustness to mislabeled training samples (using the context information in the learning phase of the algorithm). Another promising approach is active learning, which allows one an interactive analysis of RS data, by driving the user to label unlabeled samples that are selected by a query function as most informative.

We can conclude that the SVM approach showed to be very promising for the classification of RS data and recent works demonstrate that SVM can be used as basis for the development of advanced techniques for solving specific RS problems or for exploiting particular properties of the RS data. However, still effort should be devoted to the development of advanced techniques that can effectively extract useful information from the rich and complex data acquired by the last generation of RS sensors. Moreover, effort is required also for applying the SVM-based approaches developed in the research activities in real-world RS problems. Indeed, at the present, the most of the real problems related to RS image classification are still solved with standard classifiers (like maximum likelihood or *k*-NN) that, even if simple, cannot guarantee the accuracy and generalization capabilities of SVMs in complex problems.

2.5 References

[1] J.A. Benediktsson, P.H. Swain, and O.K. Ersoy, “Neural Network Approaches Versus Statistical Methods in Classification of Multisource Remote Sensing Data,” IEEE Transactions on Geoscience and Remote Sensing, vol. GE-28, no. 4, pp. 540-552, July 1990.

[2] P.D. Heermann, N. Khazenie, **“**Classification of multispectral remote sensing data using a back-propagation neural network”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 30, no. 1, pp. 81 - 88, Jan. 1992.

[3] H. Bischof, W. Schneider, A.J. Pinz, “Multispectral classification of Landsat-images using neural networks”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 30, no. 3, pp. 482 - 490, May 1992.

[4] L. Bruzzone, D.F. Prieto, “A technique for the selection of kernel-function parameters in RBF neural networks for classification of remote-sensing images”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 2, Part 2, pp. 1179 - 1184, March 1999.

[5] L. Bruzzone, M. Marconcini, U. Wegmüller, and A. Wiesmann, “An Advanced System for the Automatic Classification of Multitemporal SAR Images”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 6, pp. 1351 - 1362, June 2004.

[6] S.B. Serpico, and F. Roli, “Classification of multisensory remote-sensing images by structured neural networks”, *IEEE Transactions on Geoscience and Remote Sensing,* vol. 33, no. 3, pp. 562-578, May 1995.

[7] L. Bruzzone, D.F. Prieto, S.B. Serpico, “A neural-statistical approach to multitemporal and multisource remote-sensing image classification”, *IEEE Transactions on Geoscience and Remote Sensing,* vol. 37, no. 3, Part 1, pp. 1350 – 1359, May 1999.

[8] J.A. Benediktsson, I. Kanellopoulos, “Classification of multisource and hyperspectral data based on decision fusion”, *IEEE Transactions on Geoscience and Remote Sensing,* vol. 37, no. 3, Part 1, pp. 1367 - 1377, May 1999.

[9] L.O. Jimenez, A. Morales-Morell, A. Creus, **“**Classification of hyperdimensional data based on feature and decision fusion approaches using projection pursuit, majority voting, and neural networks”, *IEEE Transactions on Geoscience and Remote Sensing,* vol. 37, no. 3, Part 1, pp. 1360 - 1366, May 1999.

[10] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[11] J.A. Gualtieri and R.F. Cromp, "Support Vector Machines for Hyperspectral Remote Sensing Classification", 27'th AIPR Workshop, *Proc. of the SPIE*, vol. 3584, pp. 221-232, 1998.

[12] J.A. Gualtieri, S.R. Chettri, R.F. Cromp, and L.F. Johnson, "Support Vector Machine Classifiers as Applied to AVIRIS Data", Summaries of the Eighth JPL Airborne Earth Science Workshop, JPL Publication 99-17, pp. 217-227, 1999, ftp://popo.jpl.nasa.gov/pub/docs/workshops/99_docs/toc.html.

[13] J. A. Gualtieri and S. Chettri, “Support vector machines for classification of hyperspectral data”, *Proc. of IEEE-IGARSS 2000*, Hawaii, pp. 813-815, 2000.

[14] F. Melgani, L. Bruzzone, “Support vector machines for classification of hyperspectral remote-sensing images”, *IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2002)*, Toronto, Canada, pp. 506-508, June 2002.

[15] G. Camps-Valls, L. Gomez-Chova, J. Calpe-Maravilla, J.D. Martin-Guerrero, E. Soria-Olivas, L. Alonso-Chorda, J. Moreno, “Robust support vector method for hyperspectral data classification and knowledge discovery”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 7, pp. 1530 - 1542 , July 2004.

[16] F. Melgani, L. Bruzzone, “Classification of hyperspectral remote-sensing images with support vector machines”, *IEEE Transactions on Geoscience and Remote Sensing,* vol. 42, no. 8, pp. 1778-1790, August 2004.

[17] G. Camps-Valls, L. Bruzzone, “Kernel-based methods for hyperspectral images classification”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 6, pp. 1351-1362, June 2005.

[18] M. Fauvel, J.A. Benediktsson, J. Chanussot, J.R. Sveinsson, “Spectral and Spatial Classification of Hyperspectral Data Using SVMs and Morphological Profiles”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 11, pp. 3804-3814, November 2008.

[19] G.M. Foody, A. Mathur, “A relative evaluation of multiclass image classification by support vector machines”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 6, pp. 1335 - 1343, June 2004.

[20] C. Huang, L.S. Davis, J.R.G. Townshend, “An assessment of support vector machines for land cover classification”, *Int. journal of Remote Sensing*, vol. 23, no. 4, pp. 725-749, 2002.

[21] G. Zhu, D. G. Blumberg, “Classification using ASTER data and SVM algorithms; The case study of Beer Sheva, Israel”, *Remote Sensing of Environment*, vol. 80, no. 2, pp. 233-240, May 2002.

[22] L. Su, M. J. Chopping, A. Rango, J. V. Martonchik, D. P. C. Peters, “Support vector machines for recognition of semi-arid vegetation types using MISR multi-angle imagery”, *Remote Sensing of Environment*, vol. 107, no. 1-2, pp. 299-311, March 2007.

[23] J. Keuchel, S. Naumann, M. Heiler, A. Siegmund, “Automatic land cover analysis for Tenerife by supervised classification using remotely sensed data”, *Remote Sensing of Environment*, vol. 86, no. 4, pp. 530-541, 2003.

[24] I. Lizarazo, “SVM-based segmentation and classification of remotely sensed data”, *int. journal of Remote Sensing*, vol. 29, no. 24, pp. 7277–7283, December 2008.

[25] B. Dixon, N. Candade, “Multispectral landuse classification using neural networks and support vector machines: one or the other, or both?”, *Int. journal of Remote Sensing*, vol. 29, no. 4, pp. 1185–1206, February 2008.

[26] H. Carrao, P. Goncalves, M. Caetano, “Contribution of multispectral and multitemporal information from MODIS images to land cover classification”, *Remote Sensing of Environment*, vol. 112, no. 3, pp. 986–997, March 2008.

[27] L. Bruzzone, L. Carlin, “A Multilevel Context-Based System for Classification of Very High Spatial Resolution Images”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 9, 2006, 2587-2600.

[28] B. Waske, J.A. Benediktsson, “Fusion of Support Vector Machines for Classification of Multisensor Data”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 12, December 2007, pp. 3858-3866.

[29] B. Waske, S. Van der Linden, “Classifying Multilevel Imagery From SAR and Optical Sensors by Decision Fusion ”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 5, pp. 1457-1466, May 2008.

[30] M. Dalponte, L. Bruzzone, and D. Gianelle, “Fusion of Hyperspectral and LIDAR Remote Sensing Data for Classification of Complex Forest Areas”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 5, pp. 1416-1427, May 2008.

[31] C. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[32] N. Cristianini and J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods, Cambridge, U.K.: University press, 1995.

[33] B. Schölkopf and A. Smola, *Learning With Kernels*, Cambridge, MA: MIT Press, online: http://www.learning-with-kernels.org, 2002.

[34] Chih-Wei Hsu; Chih-Jen Lin, “A comparison of methods for multiclass support vector machines”, *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415 - 425, March 2002.

[35] Lili Cheng, Jianpei Zhang, Jing Yang, Jun Ma, “An Improved Hierarchical Multi-Class Support Vector Machine with Binary Tree Architecture”, *International Conference on Internet Computing in Science and Engineering*, 2008.

[36] G. F. Hughes, “On the mean accuracy of statistical pattern recognition,” *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 55–63, January 1968.

[37] L. Bruzzone, M. Chi, M. Marconcini, “Transductive SVM for Semisupervised Classification of Hyperspectral Data,” *IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2005)*, Seoul, Korea, July 2005.

[38] L. Bruzzone, M. Chi, M. Marconcini, “A Novel Transductive SVM for the Semisupervised Classification of Remote-Sensing Images”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 11, pp. 3363-3373, 2006.

[39] M. Chi, L. Bruzzone, “Semi-supervised Classification of Hyperspectral Images by SVMs Optimized in the Primal”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 6, Part 2, pp. 1870-1880, 2007.

[40] L. Gomez-Chova, G. Camps-Valls, J. Munoz-Mari, J. Calpe, “Semisupervised Image Classification With Laplacian Support Vector Machines”, *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 3, pp. 336-340, July 2008.

[41] M. Belkin, P. Niyogi, and V. Sindhwani, “Manifold regularization: A geometric framework for learning from labeled and unlabeled examples,” J. Mach. Learn. Res., vol. 7, pp. 2399–2434, Nov. 2006.

[42] G. Camps-Valls, T. Bandos Marsheva, D. Zhou, “Semi-Supervised Graph-Based Hyperspectral Image Classification”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 10, pp. 3044-3054, October 2007.

[43] L. Bruzzone, M. Chi, M. Marconcini, “Semisupervised Support Vector Machines for Classification of Hyperspectral Remote Sensing Images”, in *Hyperspectral Data Exploitation: Theory and Applications*, Ed: C-I. Chang**,** *John Wiley & Sons, Inc.*, 2007, chapter 11, pp. 275-311.

[44] A.A. Farag, R. M. Mohamed, A. El-Baz, “A unified framework for MAP estimation in remote sensing image segmentation,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 7, pp. 1617-1634, July 2005.

[45] G. Camp-Valls, L. Gomez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J Calpe-Maravilla, “Composite Kernels for Hyperspectral Image Classification” *IEEE Geoscience and Remote Sensing Letters*, vol. 3, no. 1, January 2006.

[46] M. Marconcini, G. Camps-Valls, L. Bruzzone, “A Composite Semisupervised SVM for Classification of Hyperspectral Images”, *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 2, pp. 234-238, 2009.

[47] L. Bruzzone, C. Persello, “A Novel Context-Sensitive Semisupervised SVM Classifier Robust to Mislabeled Training Samples”, *IEEE Transactions on Geoscience and Remote Sensing*, 2009, in press.

[48] L. Bruzzone, M. Marconcini, “Toward an Automatic Updating of Land-Cover Maps by a Domain Adapatation SVM Classifier and a Circular Validation Strategy”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 4, pp. 1108-1122, 2009.

[49] P. Mitra, B. U. Shankar, S. K. Pal, “Segmentation of multispectral remote sensing images using active support vector machines”, *Pattern Recognition Letters*, vol. 25, pp. 1067-1074, 2004.

[50] D. Tuia, F. Ratle, F. Pacifici, A. Pozdnoukhov, M. Kanevski, F. Del Frate, D. Solimini, W. J. Emery, “Active Learning of Very-High Resolution Optical Imagery with SVM: Entropy vs Margin Sampling”, in Proc. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2008)*, Boston, USA, pp. IV-73 - IV-76, July 2008.

[51] G. Camps-Valls and L. Bruzzone, “Kernel Methods for Remote Sensing Data Analysis”, John Wiley & Sons, Inc., 2009, in press.

Chapter 3

3. A Novel Approach to the Selection of Spatially Invariant Features for the Classification of Hyperspectral Images with Improved Generalization Capability

This chapter presents a novel approach to feature selection for the classification of hyperspectral images. The proposed approach aims at selecting a subset of the original set of features that exhibits at the same time high capability to discriminate among the considered classes and high invariance in the spatial domain of the investigated scene. This approach results in a more robust classification system with improved generalization properties with respect to standard feature-selection methods. The feature selection is accomplished by defining a multiobjective criterion function made up of two terms: 1) a term that measures the class separability and 2) a term that evaluates the spatial invariance of the selected features. In order to assess the spatial invariance of the feature subset, we propose both a supervised method (which assumes that training samples acquired in two or more spatially disjoint areas are available) and a semisupervised method (which requires only a standard training set acquired in a single area of the scene and takes advantage of unlabeled samples selected in portions of the scene spatially disjoint from the training set). The choice for the supervised or semisupervised method depends on the available reference data. The multiobjective problem is solved by an evolutionary algorithm that estimates the set of Pareto-optimal solutions. Experiments carried out on a hyperspectral image acquired by the Hyperion sensor on a complex area confirmed the effectiveness of the proposed approach.

3.1 Introduction

Hyperspectral remote sensing images, which are characterized by a dense sampling of the spectral signature of the different land-cover types, represent a very rich source of information

This chapter was published on the *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 9, pp. 3180-3191.2005, September 2009. Title: “A Novel Approach to the Selection of Spatially Invariant Features for the Classification of Hyperspectral Images With Improved Generalization Capability”. Authors: L. Bruzzone, C. Persello.

for the analysis and automatic recognition of the land-cover classes. However, supervised classification of hyperspectral images is a very complex methodological problem due to many different issues [1]-[5]: 1) the small value of the ratio between the number of training samples and the number of available spectral channels (and thus of classifier parameters), which results in the Hughes phenomenon [6]; 2) the high correlation among training patterns taken from the same area, which violates the required assumption of independence of samples included in the training set (thus reducing the information conveyed to the classification algorithm by the considered samples); and 3) the nonstationary behavior of the spectral signatures of land-cover classes in the spatial domain of the scene, which is due to physical factors related to ground (e.g., different soil moisture or composition), vegetation, and atmospheric conditions. All the aforementioned issues result in decreasing the robustness, the generalization capability, and the overall accuracy of classification systems used to generate the land-cover maps.

In order to address the abovementioned problems, in the recent literature different promising approaches have been proposed for hyperspectral image classification (as presented in the previous chapter). Among the others, we recall: 1) the use of supervised kernel methods (and in particular of Support Vector Machines), which are intrinsically robust to the Hughes phenomenon [1],[2]; 2) the use of semisupervised learning methods that take into account both labeled and unlabeled samples in the learning of the classifier [3]; and 3) the joint use of kernel methods and semisupervised techniques [4],[5]. On the one hand, SVMs are supervised classifiers that result in augmented generalization capability with respect to other classification methods thanks to the structural risk minimization principle, which allows one to effectively control the tradeoff between the empirical risk and the generalization property. On the other hand, semisupervised approaches can increase the capability of classification algorithms to derive discrimination rules that better fit with the nonstationary behavior of features in the hyperspectral image under investigation, by considering also the information of unlabeled samples. These classification methods proved to be quite effective in mitigating some of the aforementioned problems. Nevertheless, the problem of the spatial variability of the features can be addressed (together with the sample size problem) at a different and complementary level, i.e., in the feature extraction and/or feature selection phase. To this purpose, the feature extraction phase should aim at deriving discriminative features that are also as stationary as possible in the spatial domain. The feature selection phase should aim at selecting a subset of the available features that satisfies the following: 1) allows the classifier to effectively discriminate the considered classes, 2) contains features that have the most invariant as possible behavior in the spatial domain. In this chapter we focus on the development of a feature-selection approach to the identification of robust and spatially invariant features. It is worth noting that, although in the literature several feature-selection algorithms have been proposed for the analysis of hyperspectral data (e.g., [9]-[12]), to the authors' knowledge, little attention has been devoted to the aforementioned problem.

The feature-selection techniques that are most widely used in remote sensing generally require the definition of a criterion function and a search strategy. The criterion function is a measure of the effectiveness of the considered subset of features, and the search strategy is an algorithm that aims at efficiently finding a solution (i.e., a subset of features) that optimizes the adopted criterion function. In standard feature-selection methods [9]-[17], the criterion functions typically adopted are statistical measures that assess the separability of the different classes on a given training set, but do not explicitly take into account the stationarity of the features (e.g., the variability of the spectral signature of the land-cover classes). This approach may result in selecting a subset of features that retain very good discrimination properties in the portion of the scene close to the training pixels (and therefore with similar behavior), but are not appropriate to model the class distributions in separate portions on the scene, which may present different spectral behavior. Considering the typical high spatial variability of the spectral signature of land cover classes in hyperspectral images, this approach can lead to an *overfitting* phenomenon in the feature-selection phase, resulting in poor generalization capabilities of the classification system. Note that we use here the term *overfitting* with an extended meaning with respect to the conventional sense, which traditionally refers to the phenomenon that occur when inductive algorithms models too closely the training data, loosing generalization capability. In this work, we observe that there is an intrinsic spatial variability of the spectral signature of classes in the hyperspectral image, and thus, we expect that the generalization ability of the system is strongly affected from this property of hyperspectral data, which is much more critical than in standard multispectral images.

In this chapter we address the aforementioned problem by proposing a novel approach to feature selection that aims at identifying a subset of features that exhibit both high discrimination ability among the considered classes and high invariance in the spatial domain of the investigated scene. This approach is implemented by defining a novel criterion function that is based on the evaluation of two terms: 1) a standard separability measure and 2) a novel invariance measure that assesses the stationarity of features in the spatial domain. The search algorithm, adopted for deriving the subsets of features that jointly optimize the two terms, is based on the optimization of a multiobjective problem for the estimation of the Pareto-optimal solutions. For the assessment of the two terms of the criterion function we propose both a supervised and a semisupervised method that can be adopted according to the amount of available reference data. The proposed approach can be integrated in the design of any system for hyperspectral image classification (e.g., based on parametric or distribution-free supervised algorithms, kernel methods, and semisupervised classification techniques) for increasing the robustness and the generalization capability of the classifier.

This chapter is organized into six sections. The next section presents the background and a brief overview on existing feature-selection algorithms for the classification of hyperspectral data. Section 3.3 presents the proposed novel approach to the selection of features for the classification of hyperspectral images, and two possible methods to implement it according to the available reference data. Section 3.4 describes the adopted hyperspectral data set and the design of the experimental analysis carried out for assessing the effectiveness of the proposed approach. Section 3.5 presents the obtained experimental results on the considered data set. Section 3.6 draws the conclusions of this chapter.

3.2 Background on feature selection in hyperspectral images

The process of feature selection aims at reducing the dimensionality of the original feature space by selecting an effective subset of the original features, while discarding the remaining measures. Note that this approach is different from feature transformation (extraction), which consists in projecting the original feature space onto a different (usually lower dimensional) feature space [9], [14], [18], [19]. In this chapter we focus our attention on feature selection, which has the important advantage to preserve the physical meaning of the selected features. Moreover, feature selection results in a more general approach than feature transformation alone by considering that the features given as input to the feature-selection module can be associated with the original spectral channels of the hyperspectral image and/or with measures that extract information from the original channels and from the spatial context of each single pixel [20], [21] (e.g. texture, wavelets, average of groups of contiguous bands, derivatives of the spectral signature, etc).

Let us formalize a general feature-selection problem for the classification of a hyperspectral image $\mathcal{I}$, where each pixel, described by a feature vector $\mathbf{x} = (x_1, x_2, ..., x_d)$ in an *d*-dimensional feature space, is to be assigned to one of *L* different classes $\Omega = \{\omega_1, \omega_2, ..., \omega_L\}$. The set Υ is made up of the *d* features in input to the feature-selection process (which can be the original channels and/or measures extracted from them). Let $P(\omega_i)$, $\omega_i \in \Omega$, be the *a priori* probabilities of the land-cover classes in the considered scene, and $p(\mathbf{x} \mid \omega_i)$ be the conditional probability density functions for the feature vector $\mathbf{x}$, given the class $\omega_i \in \Omega$. Let us further assume that a training set $T = \{\mathcal{X}, \mathcal{Y}\}$ made up of *N* pairs $(\mathbf{x}_i, y_i)$ is available, where $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, $\mathbf{x}_i \in \mathbb{R}^d$, $\forall i = 1, 2, ..., N$, is a subset of $\mathcal{I}$ and $\mathcal{Y} = \{y_1, y_2, ..., y_N\}$, $y_i \in \Omega$, $\forall i = 1, 2, ..., N$ is the corresponding set of class labels. The aim of the feature-selection process is to select the most effective subset $\boldsymbol{\theta}^* \subset \Upsilon$ of *l* features (with *l* < *d*), according to a criterion function and a search strategy. This can be obtained according to different algorithms that broadly fall into three categories [22]: 1) the *filter* model; 2) the *wrapper* model; and 3) the *hybrid* model. The filter model is based on general characteristics of the considered data and filters out the most irrelevant features without involving the classification algorithm. Usually this is accomplished according to a measure that assesses the separability among classes. The wrapper model depends on a particular classification algorithm and exploits the classifier performance as the criterion function. It searches for a subset of features that optimize the accuracy of the adopted inductive algorithm, but it is generally computationally more expensive than the filter model. The hybrid model takes advantage of the aforementioned two models by exploiting their different evaluation criteria in different search stages. It uses a criterion function that depends on the available data to identify the subset of candidate solutions for a given cardinality *l* and then exploits the classification algorithm to select the final best subset. In the next subsections, we focus our literature analysis on the filter methods and only on the background concepts that are relevant for the developed technique.

3.2.1 Criterion functions

In standard filter approaches to feature selection, the typically adopted criterion functions are based on statistical distance measures that assess the separability among class distributions $p(\mathbf{x} \mid \omega_i)$, $\forall \omega_i \in \Omega$, on the basis of the available training set *T*. Statistical distance measures are usually adopted as they represent practical criteria to easily approximate the Bayes error. Commonly adopted measures to evaluate the separability between the distributions of two classes ω_i and ω_j, are [9], [14]:

$$\text{Divergence: } Div_{ij}(\boldsymbol{\theta}) = \int_{\mathbf{x}} \{p(\mathbf{x}|\omega_i) - p(\mathbf{x}|\omega_j)\} \ln \frac{p(\mathbf{x}|\omega_i)}{p(\mathbf{x}|\omega_j)} d\mathbf{x} \tag{3.1}$$

$$\text{Bhattacharyya distance: } B_{ij}(\boldsymbol{\theta}) = -\ln \left\{ \int_{\mathbf{x}} \sqrt{p(\mathbf{x} \mid \omega_i) p(\mathbf{x} \mid \omega_j)} d\mathbf{x} \right\} \tag{3.2}$$

$$\text{Jeffries-Matusita distance: } JM_{ij}(\boldsymbol{\theta}) = \left\{ \int_{\mathbf{x}} \left[\sqrt{p(\mathbf{x}|\omega_i)} - \sqrt{p(\mathbf{x}|\omega_j)} \right]^2 d\mathbf{x} \right\}^{1/2}. \qquad (3.3)$$

The JM distance can be rewritten according to the Bhattacharyya distance B_{ij} :

$$JM_{ij}(\boldsymbol{\theta}) = \sqrt{2\{1 - \exp[-B_{ij}(\boldsymbol{\theta})]\}} \qquad (3.4)$$

In multispectral and hyperspectral remote sensing images, the distributions of classes $p(\mathbf{x}\,|\,\omega_i)\,, \omega_i \in \Omega$ are usually modeled with Gaussian functions with mean vectors μ_i and covariance matrixes $\boldsymbol{\Sigma}_i$. Under this assumption, we can write:

$$Div_{ij}(\boldsymbol{\theta}) = \frac{1}{2} Tr\left\{ (\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_j)(\boldsymbol{\Sigma}_j^{-1} - \boldsymbol{\Sigma}_i^{-1}) \right\} + \frac{1}{2} Tr\left\{ (\boldsymbol{\Sigma}_i^{-1} - \boldsymbol{\Sigma}_j^{-1})(\mu_i - \mu_j)(\mu_i - \mu_j)^T \right\} \qquad (3.5)$$

$$B_{ij}(\boldsymbol{\theta}) = \frac{1}{8}(\mu_i - \mu_j)^T \left(\frac{\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j}{2} \right)^{-1} (\mu_i - \mu_j) + \frac{1}{2} \ln\left(\frac{1}{2} \frac{|\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j|}{\sqrt{|\boldsymbol{\Sigma}_i||\boldsymbol{\Sigma}_j|}} \right) \qquad (3.6)$$

where $Tr\{\cdot\}$ is the trace of a matrix. An important drawback of the divergence is that its value quadratically increases with respect to the separation between the mean vectors of the classes distributions. This behavior does not reflect the classification accuracy behavior, which asymptotically tends to one when the class distributions are perfectly separated. On the contrary, the JM distance exhibits a behavior that saturates when the separability between the two considered classes increases. For this reason the JM distance is generally preferred to either the divergence or the Bhattacharyya distance.

The previously described measures evaluate the statistical distance between a pair of class distributions. In order to extend the separability measures to multi-class problems, a usually adopted separability indicator is obtained by computing the average distance among all pair wise distances. Thus, a multiclass separability measure can be defined as:

$$\Delta(\boldsymbol{\theta}) = \sum_{i=1}^{L} \sum_{j>i}^{L} P(\omega_i) P(\omega_j) S_{ij}(\boldsymbol{\theta}) \qquad (3.7)$$

where $S_{ij}(\boldsymbol{\theta})$ is a statistical distance measure (e.g., Bhattacharyya distance, Divergence, and JM distance) between the distributions $p(\mathbf{x}\,|\,\omega_i)$ and $p(\mathbf{x}\,|\,\omega_j)$ of the two classes ω_i and ω_j, respectively, and $P(\omega_i)$, $P(\omega_j)$ are the prior probabilities of the classes ω_i and ω_j in the considered scene.

Other measures adopted for feature selection are based on scatter matrices that allow one characterizing the variance within classes and between classes [14]. Using these measures, the canonical analysis aims at maximizing the ratio between among-class variance and within-class variance, resulting in the selection of features that simultaneously exhibit both requirements, i.e., high among-class variance and low within-class variance. Another example of indicator that can be adopted as criterion function is the mutual information, which measures the mutual dependence of two random variables. In the context of feature selection, the mutual information can be used to assess the capability of the considered feature vector $\mathbf{x}_i \in \boldsymbol{\theta}$ to predict the correct class label $y_i \in \Omega$, $\forall i = 1, 2, ..., l$. To this purpose, a definition of the mutual information that considers the discrete nature of y should be adopted (for deeper insight on feature selection based on mutual information, we refer the reader to [23], [24]).

3.2.2 Search strategies

In order to select the final subset of features that optimizes the adopted criterion function, a search strategy is needed. The search strategy generates possible solutions of the feature-selection algorithm and compares them by applying the criterion function as a measure of the effectiveness of each solution. An exhaustive search for the optimal solution involves the evaluation and comparison of the criterion function for all $\binom{l}{d}$ possible combination of features. This is an intractable problem from a computational point of view, even for low numbers of features [17]. The *branch and bound* method proposed by Naredra and Fukunaga [14], [15] is a widely used approach to compute the globally optimum solution for monotonic criterion function without explicitly exploring all possible combinations of features. Nevertheless, the computational saving is not sufficient for treating problems with hundreds of features. Therefore, in the case of feature selection for hyperspectral data classification, suboptimal approaches should be adopted. Several suboptimal search strategies have been proposed in the literature. The simplest suboptimal search strategies are the *sequential forward selection* (SFS) and the *sequential backward selection* (SBS) techniques [16], [17]. A serious drawback of both algorithms is that they do not allow backtracking. In the case of the SFS algorithm, once the features have been selected, they cannot be discarded. Similarly, in the case of the SBS search technique, once the features have been discarded, they cannot be added again to the subset of selected features. Two effective sequential search methods are those proposed by Pudil et al. [16], namely, the *sequential forward floating selection* (SFFS) method and the *sequential backward floating selection* (SBFS) method. They improve the standard SFS and SBS techniques by dynamically changing the number of features included (SFFS) or removed (SBFS) to the subset of selected features at each step, thus allowing the reconsideration of the features included or removed at the previous steps. Other effective strategies are those proposed in [12], where two search algorithms are presented (i.e., the *steepest ascent* and the *fast constrained search*), which are based on the formalization of the feature-selection problem in the framework of a discrete optimization problem in an adequately defined binary multidimensional space.

An alternative approach to the exploration of the feature space that is relevant to this chapter, is that based on genetic algorithms (GAs), which application to feature-selection problems was proposed in [25]. Genetic algorithms exploit an analogy with biology, in which a group of solutions, encoded as *chromosomes*, evolve via natural selection [26]. A standard GA starts by randomly creating an initial population (with a predefined size). Solutions are then combined via a crossover operator to produce offspring, thus expanding the current population. The individuals in the population are evaluated according to the criterion function and the individuals that less fit such a function are discarded to return the population to its original size. A mutation operator is generally applied in order to increase individuals' variations. The processes of crossover, evaluation, and selection are repeated for a predetermined number of generations (if no other stop criterion is met before) in order to reach a satisfactory solution. Several papers confirmed the effectiveness of genetic algorithms for standard feature-selection approaches (e.g., [27]-[29]), also for hyperdimensional feature space. Moreover, as it will be explained later, GAs become particularly relevant for this work as they are effective when the criterion function involves multiple concurrent terms, and therefore a multiobjective problem has to be optimized in order to estimate the Pareto-optimal solutions [30], [31].

3.3 Proposed feature selection approach

The main idea and novelty of the approach that we propose in this chapter is to explicitly consider in the criterion function of the feature-selection process the spatial variability of the features (e.g., of the spectral signatures) on each land-cover class in the investigated scene together with their discrimination capability. This results in the possibility to select a subset of features that exhibits both high capability to discriminate among different classes and high invariance in the spatial domain. The resulting subset of selected features implicitly improves the generalization capability in the classification process, which results in augmented robustness and accuracy in the classification of hyperspectral images with respect to feature subsets selected with standard methods. This property is particularly relevant when the considered scene is extended over large geographical areas and/or presents considerable intra-class variability of the spectral signatures.

From a formal viewpoint, the aim of the proposed approach is to select the subset $\boldsymbol{\theta}^* \subset \Upsilon$ of l features (with $l < d$) that optimizes a novel criterion function made up of two measures that characterize the following: 1) the capability of the subset of features to discriminate among the considered classes in Ω and 2) the spatial invariance (stationary behavior) of the selected features. The first measure can be evaluated with standard statistical separability indices (as described in the previous section). Whereas, the spatial invariance property is evaluated according to a novel invariance measure that represents an important contribution of this work. In particular we propose two possible methods to evaluate the invariance of a subset of features: 1) a supervised method and 2) a semisupervised method. The supervised method relies on the assumption that the available training set T is made up of two subsets of labeled patterns T_1 and T_2 (such that $T_1 \cup T_2 = T$ and $T_1 \cap T_2 = \varnothing$) collected on disjoint (separate) areas on the ground. This property of the training set is exploited for assessing the spatial variability of the spectral signatures of the land-cover classes. We successively relax this hypothesis by proposing a semisupervised method that does not require the availability of a training subset T_2 spatially disjoint from T_1 (only a standard training set $T \equiv T_1$ acquired in a single area of the scene is needed) and takes advantage of unlabeled samples. This second method is based on an estimation of the distributions of classes in portions of the image separate from T, which is carried out by exploiting the information captured from unlabeled pixels. The final subset of features is selected by jointly optimizing the two concurrent terms of the criterion function. This is done by defining a proper search strategy based on the optimization of a multiobjective problem for deriving the subsets of features that exhibits the best trade-off between the two concurrent objectives.

In the following subsections we present the proposed supervised and semisupervised methods for the evaluation of the criterion function. Then we describe the proposed multiobjective search strategy for deriving the final subsets of features that exhibits both the aforementioned properties (which can be assessed with either the supervised or the semisupervised method depending on the available reference data).

3.3.1 Supervised formulation of the proposed criterion function

Let us first assume the availability of two subsets of labeled patterns T_1 and T_2 collected on disjoint areas on the ground (thus, representing two different realizations of the class distributions). Under this assumption, we can define a novel criterion function that is based on two different terms: a) a term that measures the class separability (discrimination term); b) a term that evaluates the spatial invariance of the investigated features (invariance term).

a) *Discrimination Term* Δ - This term is based on a standard feature-selection criterion function. In the proposed system we adopt the definition given in (3.7) where the term $\Delta(\boldsymbol{\theta})$ evaluates the average measure of distance between all couples of class distributions $p(\mathbf{x}\,|\,\omega_i)$ and $p(\mathbf{x}\,|\,\omega_j)$, $\forall \omega_i, \omega_j \in \Omega$ and $i < j$. This term depends on the selected subset $\boldsymbol{\theta}$ of features, and the subset of *l* features $\boldsymbol{\theta}^*$ that maximizes this distance results in the best potential for discriminating land-cover classes in the area modeled by the training samples. It is important to note that the evaluation of the above term is usually performed by assuming Gaussian distributions of classes for calculating the statistical distance $S_{ij}(\boldsymbol{\theta})$. Under this assumption, also in presence of two disjoint training sets, it is preferable to evaluate the discrimination term by considering only one subset of the training set (T_1 or T_2). This can be explained by considering that mixing up the two available training subset T_1 and T_2 would result in mixing together two different realizations of the feature distributions, which, from a theoretical perspective, can not be correctly modeled with Gaussian (mono-modal) distributions.

b) *Invariance Term* P - In order to introduce the invariance term let us first consider Fig. 3.1. This figure shows a qualitative example in a 2-dimensional feature space of two subsets of features that exhibit different behavior of the samples extracted from different portions of a scene. The features of Fig. 3.1(a) present good capability to separate the class clusters and also exhibit high invariance on the two considered training sets. These properties allow the supervised algorithm to derive a robust classification rule, resulting in the capability to accurately classify samples that can be localized in both the areas from which the samples of T_1 and T_2 are extracted. On the contrary, the features adopted in Fig. 3.1(b) exhibit good separability properties but low invariance. This feature subset leads the supervised learner to derive a classification rule that is not robust, resulting in poor classification accuracy in spatially disjoint areas.

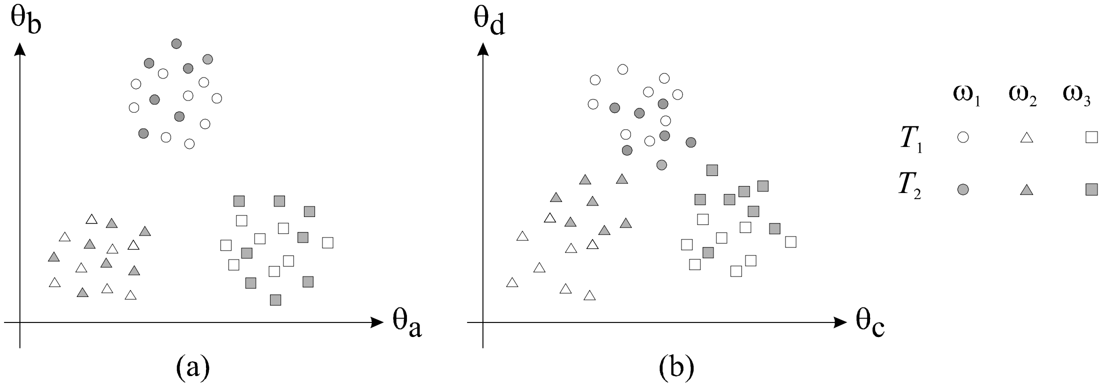


Fig. 3.1 - Examples of feature subsets with different invariant (stationary) behaviors on two disjoints set T_1 and T_2. (a) Feature subset that exhibits high separability and high invariance properties. (b) Feature subset with high separability on T_1 but high variability between T_1 and T_2.

The different behavior between the feature subsets in Fig. 3.1(a) and Fig. 3.1(b) can be modeled by considering the distance between the clusters that refer to the same land-cover class in the two disjoint training sets T_1 and T_2. Thus, we can introduce a novel term to explicitly measure the invariance (stationary behavior) of features on each class in the investigated image. It can be defined as:

$$\mathrm{P}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{L} P^{T_1}(\omega_i)P^{T_2}(\omega_i)S_{ii}^{T_1T_2}(\boldsymbol{\theta}) \tag{3.8}$$

where $S_{ii}^{T_1T_2}$ is a statistical distance measure between the distributions $p^{T_r}(\mathbf{x}\,|\,\omega_i)$, $r=1,2$ of the class ω_i computed on *T₁* and *T₂*, and $P^{T_r}(\omega_i)$ represents the prior probability of the class ω_i in T_r, $r=1,2$. This term evaluates the average distance between the distributions of the same class in different portions of the scene (i.e., on the two disjoint subsets of the training set). Unlike for $\Delta(\boldsymbol{\theta})$, we expect that a good (i.e., robust) subset of features should minimize the value of $\mathrm{P}(\boldsymbol{\theta})$. The computation of $\mathrm{P}(\boldsymbol{\theta})$ can be easily extended to more than two training subsets if labeled data collected on more than two disjoint regions are available. In the general case, when *R* spatially disjoints training sets are available, the invariance term can be defined as follows:

$$\mathrm{P}(\boldsymbol{\theta}) = \frac{1}{R}\sum_{a=1}^{R}\sum_{b>a}^{R}\sum_{i=1}^{L} P^{T_a}(\omega_i)P^{T_b}(\omega_i)S_{ii}^{T_aT_b}(\boldsymbol{\theta}) \tag{3.9}$$

The process of selection of features that jointly optimize the discrimination term $\Delta(\boldsymbol{\theta})$ and the invariance term $\mathrm{P}(\boldsymbol{\theta})$ will be described in section 3.3.3.

3.3.2 Semisupervised formulation of the criterion function (invariance term estimation)

The collection of labeled training samples on two (or more) spatially-disjoint areas from the site under investigation can be difficult and/or very expensive. This may compromise the applicability of the proposed supervised method in some real classification applications. In order to overcome this possible problem, in this section we propose a semisupervised technique to estimate the invariance term defined in (3.8), which does not require the availability of a disjoint training subset T_2. Here, we only assume that a training set T_1 is available and we consider a set of unlabeled pixels $U=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_u\}\in\mathcal{I}$ (subset of the original image $\mathcal{I}$) that should satisfy two requirements: 1) *U* contains samples of all the considered classes, and 2) samples in *U* should be taken from portions of the scene separated from those on which the training samples T_1 are collected. The set *U* can be defined: 1) by manually selecting clusters of pixels on a portion of the considered scene; 2) by randomly sub-sampling a set of pixels; or 3) by considering the whole image $\mathcal{I}$. It is worth noting that, in the proposed algorithm, the labels of classes are not required. We only assume that the unlabeled samples are collected according to a strategy that can implicitly consider all classes present in the scene.

The method is based on the semisupervised estimation of the terms $P^U(\omega_i)$ and $p^U(\mathbf{x}\,|\,\omega_i)$, $\omega_i\in\boldsymbol{\Omega}$, which, in this case, characterize the prior probabilities and the conditional probability density functions in the disjoint area corresponding to the pixels in *U*, respectively. The distribution of the samples in *U* can be described by the following mixture model:

$$p^U(\mathbf{x}) = \sum_{i=1}^{L} P^U(\omega_i)p^U(\mathbf{x}\,|\,\omega_i). \tag{3.10}$$

We assume that $P^U(\omega_i)$ and $p^U(\mathbf{x}\,|\,\omega_i)$ are not known, while $p^U(\mathbf{x})$ is given from the data distribution. However, despite the expected variability, for each class $\omega_i\in\boldsymbol{\Omega}$, the initial values of both the prior probability $P^U(\omega_i)$ and the conditional density function $p^U(\mathbf{x}\,|\,\omega_i)$ can be roughly approximated by the prior and the conditional density function in T_1, i.e.,

$$P^{U,0}(\omega_i) = P^{T_1}(\omega_i); \qquad p^{U,0}(\mathbf{x}\,|\,\omega_i) = p^{T_1}(\mathbf{x}\,|\,\omega_i). \tag{3.11}$$

The problem can be addressed by estimating the parameters vector $\mathbf{J} = [P^U(\omega_i), \delta_i]_{i=1}^{L}$, where each component δ_i represents the vector of parameters that characterize the density function $p^U(\mathbf{x}\,|\,\omega_i)$, which, given its dependence from δ_i, can be rewritten as $p^U(\mathbf{x}\,|\,\omega_i, \delta_i)$. The components of $\mathbf{J}$ can be estimated by maximizing the pseudo log-likelihood function $L[p^U(\mathbf{x})]$defined as

$$L[p^U(\mathbf{x})\,|\,\mathbf{J}] = \sum_{j=1}^{l} \log\left\{\sum_{i=1}^{L} P^U(\omega_i\,|\,\mathbf{J})\, p^U(\mathbf{x}\,|\,\omega_i, \mathbf{J})\right\}. \tag{3.12}$$

The maximization of the log-likelihood function can be obtained with the expectation maximization (EM) algorithm [32]. The EM algorithm consists of two main steps: an expectation step and a maximization step. The two steps are iterated, so that the value of the log-likelihood function $L[p^U(\mathbf{x})]$ increases at each iteration, until a local maximum is reached. For simplicity, let us consider that all the classes $\omega_i \in \Omega$ are Gaussian distributed. Under this assumption the density function associated with each class ω_i can be completely described by the mean vector μ_i^U and the covariance matrix Σ_i^U, $i = 1,...,L$. Therefore the parameters vector to be estimated becomes:

$$\mathbf{J} = [P^U(\omega_i), \mu_i^U, \Sigma_i^U]_{i=1}^{L}. \tag{3.13}$$

It can be proven that the equations to be used at iteration s+1 for estimating the statistical terms associated with a generic class ω_i are the following [3], [32], [33] :

$$P^{U,s+1}(\omega_i) = \frac{1}{l}\sum_{\mathbf{x}_j \in U} \frac{P^{U,s}(\omega_i)\, p^{U,s}(\mathbf{x}_j\,|\,\omega_i)}{p^{U,s}(\mathbf{x}_j)} \tag{3.14}$$

$$[\mu_i^U]^{s+1} = \frac{\displaystyle\sum_{\mathbf{x}_j \in U} \frac{P^{U,s}(\omega_i)\, p^{U,s}(\mathbf{x}_j\,|\,\omega_i)}{p^{U,s}(\mathbf{x}_j)}\mathbf{x}_j}{\displaystyle\sum_{\mathbf{x}_j \in U} \frac{P^{U,s}(\omega_i)\, p^{U,s}(\mathbf{x}_j\,|\,\omega_i)}{p^{U,s}(\mathbf{x}_j)}} \tag{3.15}$$

$$[\Sigma_i^U]^{s+1} = \frac{\displaystyle\sum_{\mathbf{x}_j \in U} \frac{P^{U,s}(\omega_i)\, p^{U,s}(\mathbf{x}_j\,|\,\omega_i)}{p^{U,s}(\mathbf{x}_j)}\left\{\mathbf{x}_j - [\mu_i^U]^{s+1}\right\}^2}{\displaystyle\sum_{\mathbf{x}_j \in U} \frac{P^{U,s}(\omega_i)\, p^{U,s}(\mathbf{x}_j\,|\,\omega_i)}{p^{U,s}(\mathbf{x}_j)}} \tag{3.16}$$

where the superscripts s and s+1 refer to the values of the parameters at the *s-th* and $s+1$-*th* iteration, respectively. The estimates of the statistical parameters that describes the classes distributions in the disjoint areas are obtained starting from the initial values of the parameters [see (3.11)] and iterating the equations (3.14)-(3.16) up to convergence. An important aspect of the EM algorithm concerns its convergence properties. It is not possible to guarantee that the algorithm will converge to the global maximum of the log-likelihood function, although convergence to a local maximum can be ensured. A detailed description of the EM algorithm is beyond the scope of this chapter, so we refer the reader to the literature for a more detailed analysis of such an algorithm and its properties [3], [32]. The final estimates obtained at convergence for each class $\omega_i \in \Omega$, i.e., $\hat{P}^U(\omega_i)$, and $\hat{p}^U(\mathbf{x} \mid \omega_i)$ (which depend on the estimated parameters $\hat{\mu}_i^U$, $\hat{\Sigma}_i^U$) can be used in place of $P^{T_2}(\omega_i)$ and $p^{T_2}(\mathbf{x} \mid \omega_i)$ to estimate the invariance term $\hat{\mathrm{P}}(\boldsymbol{\theta})$ for each subset of features $\boldsymbol{\theta}$ considered. Thus, the semisupervised estimation of the invariance term becomes:

$$\hat{\mathrm{P}}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{L} P^{T_1}(\omega_i)\hat{P}^U(\omega_i)\hat{S}_{ii}^{T_1U}(\boldsymbol{\theta}). \tag{3.17}$$

The discrimination term $\Delta(\boldsymbol{\theta})$ can be calculated as in (3.7) with no difference with respect to the supervised method.

It is worth noting that, depending on the adopted set U of unlabeled pixels, the estimation of the prior probabilities and the class conditional densities can reflect with different degree of accuracy the true values. In particular, the estimation of the elements of the covariance matrices $\hat{\Sigma}_i^U$, $i = 1,\ldots,L$ may become critical in some cases when the number of classes is high. Thus, in these cases, since small fluctuations in the accuracy of the estimation of the covariance terms $\hat{\Sigma}_i^U$, $i = 1,\ldots,L$ can strongly affect the invariance term values, the estimation of the invariance term can be simplified: 1) by assuming that the covariance matrix is diagonal, 2) by considering only the first-order statistical moment (thus neglecting the second-order moments) for the evaluation of the statistical distance $\hat{S}_{ii}^{T_1U}(\boldsymbol{\theta})$.

3.3.3 Proposed multiobjective search strategy

Given the proposed criterion function that is made up of the discrimination term $\Delta(\boldsymbol{\theta})$ and invariance term $\mathrm{P}(\boldsymbol{\theta})$ (which, depending on the available reference data, can be evaluated with the supervised or the unsupervised methods as described in the two previous subsections), we address now the problem of defining a search strategy to select the subset (or the subsets) of features that jointly optimizes the two defined measures. To this purpose, one can define a global optimization function as

$$V(\boldsymbol{\theta}) = \Delta(\boldsymbol{\theta}) + K \cdot f\left[\mathrm{P}(\boldsymbol{\theta})\right] \tag{3.18}$$

where K tunes the tradeoff between discrimination ability and invariance of the selected subset of features, and f is monotonic decreasing function of $\mathrm{P}(\boldsymbol{\theta})$. The subset $\boldsymbol{\theta}^*$ of l features for which $V(\boldsymbol{\theta})$ has the maximum value represents the solution to the considered problem.

Nevertheless, the aforementioned formulation of the problem has two drawbacks: 1) the obtained criterion function is not monotonic (and thus effective search algorithms based on this property cannot be used), and 2) the definition of f and K (which should be carried out empirically) affects significantly the final result. To overcome these drawbacks, we modeled this problem as a multiobjective minimization problem, where the multiobjective function $\mathbf{g}(\boldsymbol{\theta})$ is made up of two different (and possibly conflicting) objectives $g_1(\boldsymbol{\theta})$ and $g_2(\boldsymbol{\theta})$, which express the discrimination ability $\Delta(\boldsymbol{\theta})$ among the considered classes and the spatial invariance $\mathrm{P}(\boldsymbol{\theta})$ of the subset of features $\boldsymbol{\theta}$, respectively. The multiobjective problem can therefore be formulated as follows:

$$\begin{gathered}\min_{|\boldsymbol{\theta}|=l}\left\{\mathbf{g}(\boldsymbol{\theta})\right\}, \\ \text{where } \mathbf{g}(\boldsymbol{\theta}) = [g_1(\boldsymbol{\theta}), g_2(\boldsymbol{\theta})] = [-\Delta(\boldsymbol{\theta}), \mathrm{P}(\boldsymbol{\theta})]\end{gathered} \tag{3.19}$$

where $|\boldsymbol{\theta}|$ is the cardinality of the subset $\boldsymbol{\theta}$, i.e., the number of features l to be selected from the d originally available. This problem is solved in order to obtain a set of Pareto-optimal solutions O^*, instead of a single optimal one. In greater detail, a solution $\boldsymbol{\theta}^*$ is said to be Pareto optimal if it is not dominated by any other solution in the search space, i.e., there is no other $\boldsymbol{\theta}$ such that $g_i(\boldsymbol{\theta}) \leq g_i(\boldsymbol{\theta}^*)$ $(\forall i = 1, 2)$ and $g_j(\boldsymbol{\theta}) < g_j(\boldsymbol{\theta}^*)$ for at least one j $(\forall j = 1, 2)$. This means that $\boldsymbol{\theta}^*$ is Pareto optimal if there exists no other subset of features $\boldsymbol{\theta}$ which would decrease an objective without simultaneously increasing the other one (Fig. 3.2 clarifies this concept with a graphical example). The set O^* of all optimal solutions is called Pareto-optimal set. The plot of the objective function of all solutions in the Pareto-optimal set is called Pareto front $PF^* = \{\mathbf{g}(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in O^*\}$. Because of the complexity of the search space, an exhaustive search of the set of optimal solution O^* is unfeasible. Thus, instead of identifying the true set of optimal solutions, we aim to estimate a set of non-dominated solutions $\hat{O}^*$ with objective values as close as possible to the Pareto front. This estimation can be achieved with different multiobjective optimization algorithms (e.g., multiobjective evolutionary algorithms).

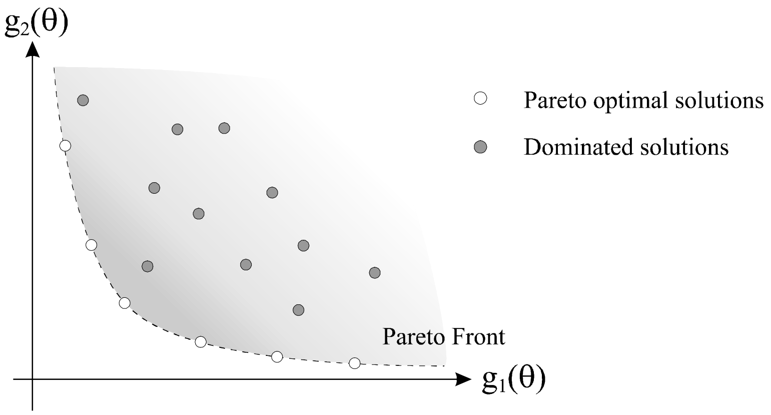


Fig. 3.2 - Example of Pareto-optimal solutions and dominated solution in a two-objective search space.

The main advantage of the multiobjective approach is that it avoids to aggregate metrics capturing multiple objectives into a single measure. On the contrary, it allows one to effectively identify different possible tradeoffs between the values of $\Delta(\boldsymbol{\theta})$ and $\mathrm{P}(\boldsymbol{\theta})$. This results in the possibility to evaluate in a more flexible way the tradeoffs between discrimination ability among classes and spatial invariance of each feature subset, and to identify the subsets of features that simultaneously exhibit both properties. In particular, we expect that the most robust subsets of features (which will results in the best generalization capability of the classification system) are represented by the solutions that are localized close to the knee of the estimated Pareto front (or the solutions closest to the origin of the search space).

3.4 Data set description and design of experiments

In order to assess the effectiveness of the presented approach (with both the proposed supervised and semisupervised methods), we carried out several experiments on a hyperspectral image acquired over an extended geographical area. We considered a data set which is increasingly used as a benchmark in the literature and consists of data acquired by the Hyperion sensor of the

EO-1 satellite in an area of the Okavango Delta, Botswana. The Hyperion sensor on EO-1 acquired the hyperspectral image with a spatial resolution of 30 m over a 7.7 km strip in 242 bands. Uncalibrated and noisy bands that cover water absorption range of the spectrum were removed, and the remaining 145 bands were given as input to the feature-selection technique. For greater details on this data set, we refer the reader to [34]. The labeled reference samples were collected on two different and spatially disjoint areas (Area 1 and Area 2), thus representing possible spatial variabilities of the spectral signatures of classes. The samples taken on the first area were partitioned into a training set T_1 and a test set TS_1 by a random sampling (these sets represent similar realizations of the spectral signatures of classes). Samples taken on the second area were used to derive a training set T_2 and test set TS_2 according to the same procedure used for the samples of the first considered area (these two sets present possible variability in class distributions with respect to the first two sets). The number of labeled reference samples for each set and class are reported in Table 3.1. After preliminary experiments carried out in order to understand the size of the subset of features that leads to the saturation of the classification accuracies, we performed different experiments (with both the supervised and the semisupervised methods) varying the size l of the selected subset of features in a range between 6 and 14 with step 2. The obtained subsets of features were used to perform the classification with a Gaussian maximum-likelihood (ML) classifier. The training of the ML classifier (estimation of Gaussian parameters for class conditional densities) was carried out using the training set T_1. We compared the classification accuracies obtained on both test sets TS_1 and *TS2* performing the feature selection with the following: 1) the proposed approach with the supervised method for the estimation of the invariance term; 2) the proposed semisupervised method for estimating the invariance term; and 3) a standard feature-selection technique that considers only the discrimination term.

Table 3.1 - Number of training (T_1 and T_2) and test (TS_1 and TS_2) patterns acquired in the two spatially disjoint areas

Class	**Number of samples**			
	Area 1		**Area 2**	
	T$_1$	**TS$_1$**	**T$_2$**	**TS$_2$**
Water	69	57	213	57
Hippo grass	81	81	83	18
Floodplain grasses1	83	75	199	52
Floodplain grasses2	74	91	169	46
Reeds1	80	88	219	50
Riparian	102	109	221	48
Firescar2	93	83	215	44
Island interior	77	77	166	37
Acacia woodlands	84	67	253	61
Acacia shrublands	101	89	202	46
Acacia grasslands	184	174	243	62
Short mopane	68	85	154	27
Mixed mopane	105	128	203	65
Exposed soil	41	48	81	14
Total	1242	1252	2621	627

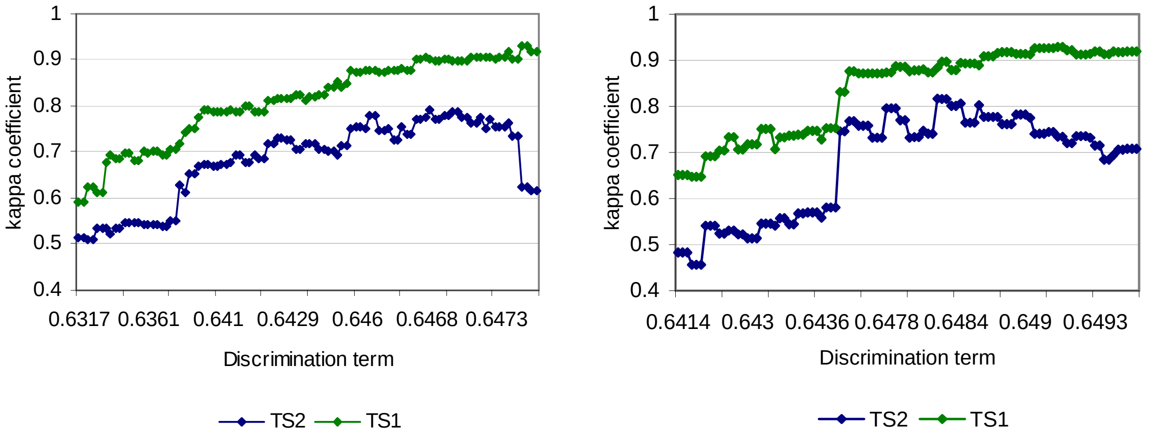
The experiments with the supervised feature-selection method were carried out by considering the training set T_1 for the evaluation of the discrimination term $\Delta(\boldsymbol{\theta})$ and both T_1 and T_2 for the evaluation of the invariance term $\mathrm{P}(\boldsymbol{\theta})$. In our implementation we adopted the JM distance (under the Gaussian assumption for the distribution of classes) as a statistical distance measure for both the considered terms. The second set of experiments was carried out with the proposed semisupervised feature-selection method. In these experiments we considered the training set T_1 for the evaluation of the discriminative term $\Delta(\boldsymbol{\theta})$, while the invariance term $\hat{\mathrm{P}}(\boldsymbol{\theta})$ was estimated from T_1 and the samples of T_2, which were used without their class label information as set U. For simplicity, we considered only the first order moment to evaluate the statistical distance $\hat{S}_{ii}^{T_1U}(\boldsymbol{\theta})$ (see discussion reported in section 3.2.1). The standard feature selection was performed by selecting the subsets of features that maximize the JM distance on the training set T_1 with a (mono-objective) genetic algorithm. Note that we did not mix up the two training set T_1 and T_2 both for training the ML classifiers and for evaluating the discrimination term, as the Gaussian approximation is no more reasonable for the two different Gaussian realizations of each class in T_1 and T_2 (see section 3.2.1).

In order to solve the defined two-objective minimization problem for the proposed methods (i.e., estimating the Pareto-optimal solutions), we implemented a modification of the "Non-Dominated Sorting in Genetic Algorithm II" (NSGA-II) [31]. The original algorithm was modified in order to avoid solutions with multiple selections of the same feature. This has been accomplished by changing the random initialization of the chromosome population and by modifying the crossover and mutation operators. In all the experiments, the population size was set equal to 100, and the maximum number of generations equal to 50. The classification was carried out using all combinations of features $\hat{\boldsymbol{\theta}}^* \in \hat{O}^*$ that lie on the estimated Pareto front, and the subset $\hat{\boldsymbol{\theta}}^*$ that resulted in the highest accuracy on the disjoint test set TS_2 was finally selected. For the mono-objective genetic algorithm we adopted the same values for both the population size and the maximum number of generations as for the multiobjective genetic algorithm.

3.5 Experimental results

3.5.1 Results with the supervised method for the estimation of the invariance term

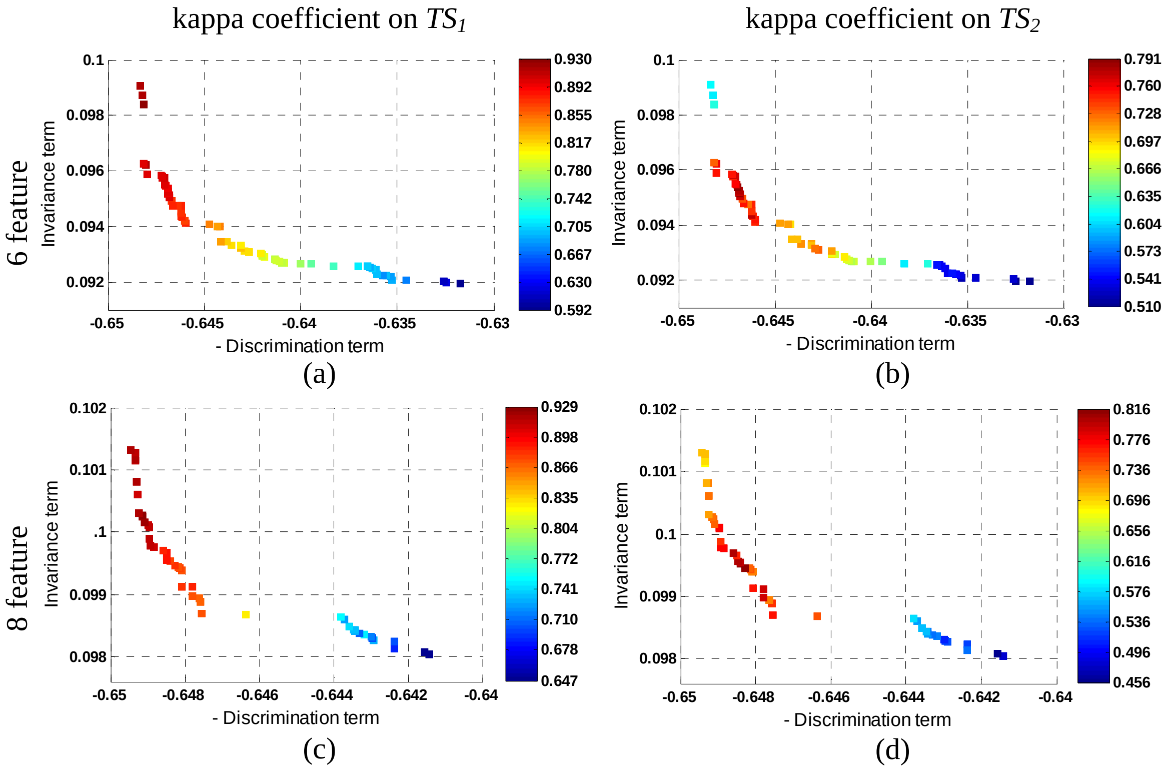
We first present the experimental results obtained with the proposed supervised method that allows us to derive important considerations about the validity of the proposed approach with respect to the standard one. In order to show the shortcomings of standard feature-selection algorithms for the classification of hyperspectral images, Fig. 3.3 plots the graphs of the accuracy obtained by the ML classifier on the adjoint (TS_1) and disjoint (TS_2) test sets versus the values of the discrimination term $\Delta(\boldsymbol{\theta})$ for different subset of features. For the reported graphs we used the solutions on the Pareto front estimated by the modified NSGA-II algorithm applied to the multiobjective minimization problem in (3.19), in the cases of six and eight features (these two cases are selected as examples; the other considered cases led to similar results). From this figure, it is possible to observe that the accuracy on TS_1 increases when the discrimination term increases, whereas the accuracy on TS_2 increases only till a certain value and then it decreases. Therefore, the simple maximization of the discrimination term (as standard approaches do) can lead to an overfitting phenomenon, which result in poor generalization capabilities, i.e., low capability to discriminate and correctly classify the land-cover classes in areas of the scene different from that associated with the collected training data. This confirms the significant variability of the spectral signature of classes in hyperspectral images.



(a) (b)

Fig. 3.3 – Behaviors of the kappa coefficients of accuracy on the test set TS_1 and TS_2 versus the values of the discrimination term $\Delta(\boldsymbol{\theta})$. Cases of (a) six and (b) eight features.

The aim of the proposed approach is to overcome this problem. Let us now consider Fig. 3.4 that depicts the Pareto fronts estimated by the proposed approach (employing the modified NSGA-II algorithm) in the cases of the selection of 6 and 8 features. This figure represents the information of the kappa coefficient of accuracy, which is obtained by the classification of the test sets TS_1 and TS_2 with the considered subset of features $\hat{\boldsymbol{\theta}}^*$, as the color of the point, according to the reported color scale bar. The diagrams in Fig. 3.4 (a)-(c) show that for the classification of TS_1, the solutions with higher discrimination capability [lower values of $-\Delta(\boldsymbol{\theta})$] result in better accuracies. This behavior reveals (as expected) that only the discrimination term is important for selecting the most effective feature subset for the classification of pixels acquired in a similar area of pixels in T_1 (in this conditions training and test patterns represent the same realization of the statistical distributions of classes). On the contrary, the diagrams in Fig. 3.4(b)-(d) show that the most accurate solutions for the classification of the spatially disjoint samples of TS_2 (which result in the highest kappa coefficient of accuracy) are located in a middle region, close to the knee of the estimated Pareto front. This confirms the importance of the invariance term, and that tradeoff solutions between the two competing objectives $\Delta(\boldsymbol{\theta})$ and $\text{P}(\boldsymbol{\theta})$ should be identified in order to select the subset of features that lead to better generalization capabilities, and thus higher classification accuracy in areas of the hyperspectral image different from the training one.



(a) (b)

(c) (d)

Fig. 3.4 – Pareto fronts estimated by the proposed approach with the supervised method. (a)-(b): 6-feature case; (c)-(d): 8-feature case. The color indicates the kappa coefficient of accuracy on (a)-(c) TS_1 and (b)-(d) TS_2 according to the reported color scale bar.

Table 3.2 reports the comparison of the classification accuracies obtained on TS_1 and TS_2 by selecting the subset of features with the proposed multiobjective supervised and semisupervised methods, as well as the standard method. From this table, it is possible to observe that the obtained accuracy on the disjoint test set TS_2 are, in general, significantly lower that those obtained on the adjoint test set TS_1, confirming the presence of consistent variability in the spatial domain of the spectral signatures of the classes. This phenomenon severely challenges the generalization capability of the classification system. Nevertheless, we can observe that for all considered cases, the proposed multiobjective feature-selection methods allowed to significantly increase the accuracy on the test set TS_2 with respect to the standard method, while the accuracy on the adjoint test set TS_1 only slightly decreased. In average, the proposed supervised method resulted in an increase of the classification accuracy on the disjoint test set of 21.3% with respect to the standard approach, slightly decreasing of 4.2% the accuracy on the adjoint test set.

The obtained results clearly confirm that the proposed approach is effective in exploiting the information of the two distinct available training sets to select subsets of robust and invariant features, which can improve the generalization capabilities of the classification system. We further observe that very few spectral channels (6-14 bands out of the originally 145 available) are sufficient for effectively representing and discriminating the considered information classes, thus significantly reducing the problems associated with the Hughes phenomenon. The computational cost of the proposed supervised method is comparable with the cost of the standard mono-objective algorithm. In our experiments, carried out on a PC mounting an Intel Pentium D processor at 3.4 GHz and a 2 Gb DDR2 RAM, the feature selection with the supervised multiobjective method took an average time of about 4 minutes, while the standard method took about 3 minutes. This is due to the fact that the evaluation of the discrimination term $\Delta(\boldsymbol{\theta})$ (which has to be computed also with standard feature-selection methods) requires a computational cost that is proportional to $L(L-1)/2$, while the introduced invariance term $\mathrm{P}(\boldsymbol{\theta})$ has a computational cost proportional to L. Therefore, the additional cost due to the evaluation of the new term becomes lesser and lesser when the number of classes increases.

3.5.2 Results with the semisupervised method for the estimation of the invariance term

Often in real applications a disjoint training set T_2 is not available to the user and the proposed supervised method can not be used. In these cases, the semisupervised approach can be adopted. It is worth noting that from the perspective of the semisupervised method, the supervised technique represents an upper bound of the accuracy and generalization ability that can be obtained (if the same samples with and without labels are considered). Thus, in this case the results presented in the previous section can be seen as the best performances that can be obtained on the considered samples.

As expected, the semisupervised method led to accuracies slightly smaller than the supervised method, but still maintained a significant improvement with respect to the traditional approach. In average, the semisupervised method increased the classification accuracy on TS_2 of 16.4% with respect to the standard feature-selection method, while decreased the accuracy on TS_1 of 3.1%. The small decrease in the performances with respect those obtained by the supervised method are due to the approximate estimation of the invariance term carried out with the EM algorithm, which can not ensure to converge to the optimal solution. However, the semisupervised method has the very important advantage to considerably increase the generalization capabilities of the classification systems with respect to the traditional approach without requiring additional reference data. The computation cost of this method is slightly higher with respect to the standard method, because of the time required by EM algorithm to perform the estimation necessary to evaluate the invariance term. In our experiments, the average time for the feature selection with the semisupervised approach was of about 60 minutes (15 times more than the supervised method).

Table 3.2 - Kappa Coefficient of Accuracies obtained by the ML classifier with the features selected by the proposed supervised and semisupervised methods, and the standard approach

Number of features	**Kappa coefficient of Accuracy on Test Set *TS_2***			**Kappa coefficient of Accuracy on Test Set *TS_1***		
	Proposed Semisup. Method	**Proposed Supervised method**	**Standard method**	**Proposed Semisup. Method**	**Proposed Supervised method**	**Standard method**
6	0.780	0.791	0.580	0.894	0.902	0.931
8	0.767	0.816	0.577	0.906	0.884	0.939
10	0.777	0.813	0.592	0.938	0.912	0.942
12	0.722	0.808	0.591	0.914	0.900	0.954
14	0.739	0.799	0.625	0.912	0.913	0.953
Average	0.757	0.805	0.593	0.913	0.902	0.944

3.6 Conclusion

In this chapter we presented a novel feature-selection approach to the classification of hyperspectral images. The proposed approach aimed at selecting subsets of features that exhibit, at the same time, high discrimination ability and high spatial invariance, improving the robustness and the generalization properties of the classification system with respect to standard techniques. The feature selection was accomplished by defining a multiobjective criterion function that considers the evaluation of both a standard separability measure and a novel term that measures the spatial invariance of the selected features. In order to assess the invariance in the scene of the feature subset we proposed both a supervised method (assuming the availability of training samples acquired in two or more spatially disjoint areas) and a semisupervised method (which requires only a standard training set acquired in a single area of the scene and exploits the information of unlabeled pixels in portions of the scene spatially disjoint from the training areas). The multiobjective problem was solved by an evolutionary algorithm for the estimation of the set of Pareto-optimal solutions.

Experimental results showed that the proposed feature-selection approach selected subsets of the original features that sharply increased the classification accuracy on disjoint test samples, while it slightly decreased the accuracy on the adjoint test set with respect to standard methods. This behavior confirms that the proposed approach results in augmented generalization capability of the classification system. In this regard, we would like to stress the importance of evaluating the accuracy on a disjoint test set, because this allows one to estimate the accuracy in the classification of the whole considered image. In particular, the proposed supervised method is effective in exploiting the information of the two available training sets, and the proposed semisupervised method can significantly increase the generalization capabilities of the classification system, without requiring additional reference data with respect to traditional feature-selection algorithms. This can be achieved at the cost of an acceptable additional computational time.

It is important to note that the proposed approach is defined in a general way, thus allowing different possible implementations. For instance, the discrimination and invariance terms can be evaluated considering statistical distance measures different from those adopted in our experimental analysis, as well as, other multiobjective optimization algorithms can be adopted as search strategy for estimating the Pareto-optimal solutions. This general definition of the approach results in the possibility to further developing the implementation that we adopted for our experimental analysis. As an example, as future developments of this work, the proposed approach could be integrated with classification algorithms different from the adopted maximum likelihood classifier, e.g., the Support Vector Machine and/or other kernel based classification techniques, for further improving the accuracy of the classification system. In addition, we think that the overall classification system can be further improved by jointly exploiting the proposed feature-selection approach and a semisupervised classification technique for a synergic and complete exploitation of the unlabeled samples information.

3.7 References

[1] F. Melgani, L. Bruzzone, "Classification of hyperspectral remote-sensing images with support vector machines", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778-1790, August 2004.

[2] G. Camps-Valls, L. Bruzzone, "Kernel-based Methods for Hyperspectral Image Classification", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 6, 1351-1362, June 2005.

[3] B. M. Shahshahani and D. A. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 5, pp. 1087–1095, September 1994.

[4] L. Bruzzone, M. Chi, M. Marconcini, "A Novel Transductive SVM for the Semisupervised Classification of Remote-Sensing Images", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 11, pp. 3363-3373, 2006.

[5] M. Chi, L. Bruzzone, "Semi-supervised Classification of Hyperspectral Images by SVMs Optimized in the Primal", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 6, Part 2, pp. 1870-1880, June 2007.

[6] G. F. Hughes, "On the mean accuracy of statistical pattern recognition," *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 55–63, January 1968.

[7] V. N. Vapnik, The Nature of Statistical Learning Theory, 2nd ed., New York: Springer, 2001.

[8] N. Cristianini and J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods, Cambridge, U.K.: University press, 1995.

[9] J. A. Richards, X. Jia, Remote Sensing Digital Image Analysis, 4th ed., Berlin, Germany: Springer-Verlag, 2006.

[10] P.W. Mausel, W.J. Kramber and J.K. Lee, "Optimum Band Selection for Supervised Classification of Multispectral Data", *Photogrammetric Engineering and Remote Sensing*, vol. 56, no. 1, pp. 55–60, January 1990.

[11] R. Archibald and G. Fann, "Feature Selection and Classification of Hyperspectral Images With Support Vector Machines", *IEEE Geoscience and Remote Sensing Letters*, vol. 4, no. 4, October 2007.

[12] S. Serpico, L. Bruzzone, "A new search Algorithm for Feature Selection in Hyperspectral Remote Sensing Images", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 7, 2001, pp. 1360-1367, July 2001.

[13] L. Bruzzone, F. Roli, S. B. Serpico, "An Extension of the Jeffreys-Matusita Distance to Multiclass Cases for Feature Selection", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 33, no. 6, 1995, pp. 1318-1321.

[14] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd ed. New York: Academic, 1990.

[15] P. M. Narendra and K. Fukunaga, “A Branch and Bound Algorithm for Feature Subset Selection”, *IEEE Transactions on Computers*, vol. C-26, pp. 917-922, September 1977.

[16] P. Pudil, J. Novovicova and J. Kittler, “Floating Search Methods for Feature Selection”, *Pattern Recognition Letter*, vol. 15, pp. 1119-1125, 1994.

[17] A. Jain and D. Zongker, “Feature selection: evaluation, application, and small sample performance,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, 1997.

[18] A. Hyvarinen and E. Oja, “Independent component analysis: Algorithms and applications,” *Neural Networks*, vol. 13, no. 4/5, pp. 411–430, May/June 2000.

[19] S. Serpico, G. Moser, “Extraction of Spectral Channels from Hyperspectral Images for Classification Purposes”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 2, 2007, pp. 484-495, February 2007.

[20] J. A. Benediktsson, M. Pesaresi, and K. Arnason, “Classification and feature extraction for remote sensing images from urban areas based on morphological transformations,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 9, pp. 1940–1949, September 2003.

[21] M. N. Do, and M. Vetteri, “Wavelet-Based Texture Retrieval Using Generalized Gaussian Density and Kullback–Leibler Distance”, *IEEE Transactions on Image Processing*, vol. 11, no. 2, pp. 146–158, February 2002.

[22] H. Liu, and L. Yu, “Toward Integrating Feature Selection Algorithms for Classification and Clustering”, *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, April 2005.

[23] H. Peng, F. Long, and C.Ding, “Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, August 2005.

[24] B. Guo, R.I. Damper, S.R. Gunn, J.D.B. Nelson, “A fast separability-based feature-selection method for high-dimensional remotely sensed image classification”, *Pattern Recognition*, vol. 41, pp. 1653-1662, November 2007.

[25] W. Siedlecki and J. Sklansky, “A note on genetic algorithms for largescale feature selection,” *Pattern Recognition Letters*, vol. 10, pp. 335–347, 1989.

[26] D. E. Goldberg, “Genetic Algorithms in Search, Optimization and Machine Learning”, Reading, MA: Addison-Wesley, 1989.

[27] F. Z. Brill, D. E. Brown, and W. N. Martin, “Fast Genetic Selection of Features for Neural Network Classifiers”, *IEEE Transactions on Neural Networks*, vol. 3, no. 2, pp. 324–328, March 1992.

[28] J. H. Yang and V. Honavar, “Feature Subset Selection Using a Genetic Algorithm,” *IEEE Intelligent Systems*, vol. 13, no. 2, pp. 44-49, 1998.

[29] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain, “Dimensionality Reduction Using Genetic Algorithms,” *IEEE Transactions on Evolutionary Computation*, vol. 4, no. 2, pp. 164-171, July 2000.

[30] H. C. Lac, D. A. Stacey, “Feature Subset Selection via Multi-Objective Genetic Algorithm”, *Proceeding of International Joint Conference on Neural Networks*, Montreal, Canada, pp. 1349-1354, July 31 – August 4, 2005.

[31] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II”, *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182-197, 2002.

[32] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. R. Statist. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[33] L. Bruzzone and D. F. Prieto, “Unsupervised Retraining of a Maximum Likelihood lassifier for the Analysis of Multitemporal Remote Sensing Images”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 2, 2001, pp. 456-460, February 2001.

[34] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, “Investigation of the Random Forest Framework for Classification of Hyperspectral Data”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, 2005, pp. 492-501.

Chapter 4

4. A Novel Context-Sensitive Semisupervised SVM Classifier Robust to Mislabeled Training Samples

This chapter presents a novel context-sensitive semisupervised Support Vector Machine (CS^{4}VM) classifier, which is aimed at addressing classification problems where the available training set is not fully reliable, i.e., some labeled samples may be associated to the wrong information class (mislabeled patterns). Unlike standard context-sensitive methods, the proposed CS^{4}VM classifier exploits the contextual information of the pixels belonging to the neighborhood system of each training sample in the learning phase to improve the robustness to possible mislabeled training patterns. This is achieved according to both the design of a semisupervised procedure and the definition of a novel contextual term in the cost function associated with the learning of the classifier. In order to assess the effectiveness of the proposed CS^{4}VM and to understand the impact of the addressed problem in real applications, we also present an extensive experimental analysis carried out on training sets that include different percentages of mislabeled patterns having different distributions on the classes. In the analysis we also study the robustness to mislabeled training patterns of some widely used supervised and semisupervised classification algorithms (i.e., conventional SVM, progressive semisupervised SVM, Maximum Likelihood, and k-Nearest Neighbor). Results obtained on a very high resolution image and on a medium resolution image confirm both the robustness and the effectiveness of the proposed CS^{4}VM with respect to standard classification algorithms and allow us to derive interesting conclusions on the effects of mislabeled patterns on different classifiers.

4.1 Introduction

The classification of remote sensing images is often performed by using supervised classification algorithms, which require the availability of labeled samples for the training of the classi-

This chapter was published on the *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2142-2154, July 2009. Title: "A Novel Context-Sensitive Semisupervised SVM Classifier Robust to Mislabeled Training Samples". Authors: L. Bruzzone, C. Persello.

fication model. All these algorithms are sharply affected from the quality of the labeled samples used for training the classifier, whose reliability is of fundamental importance for an adequate learning of the properties of the investigated scene (and thus for obtaining accurate classification maps). In supervised classification, the implicit assumption is that all labels associated with training patterns are correct. Unfortunately, in many real cases, this assumption does not hold and small amounts of training samples are associated with a wrong information class due to errors occurred in the phase of collection of labeled samples. Labeled samples can be derived by the following: 1) *in situ* ground truth surveys; 2) analysis of reliable reference maps; or 3) image photointerpretation. In all these cases, mislabeling errors are possible. During the ground truth surveys, mislabeling errors may occur due to imprecise geo-localization of the positioning system; this leads to the association of the identified land-cover label with a wrong geographic coordinate, and thus with the wrong pixel (or region of interest) in the remotely sensed image. Similar errors may occur if the image to be classified is not precisely georeferenced. When reference maps are used for extracting label information, possible errors present in the maps propagate to the training set. The case of image photointerpretation is also critical, as errors of the human operator may occur, leading to a mislabeling of the corresponding pixels or regions.

Mislabeled patterns bring distort (wrong) information to the classifier (in this thesis we call them *noisy* patterns). The effect of noisy patterns in the learning phase of a supervised classifier is to introduce a bias in the definition of the decision regions, thus decreasing the accuracy of the final classification map. We can expect two different situations with respect to the distribution of noisy samples in the training set: 1) mislabeled samples may be uniformly distributed over all considered classes, or 2) mislabeled patterns can specifically affect one or a subset of the classes of the considered classification problem. The two different situations result in a different impact on the learning phase of the classification algorithms. Let us analyze the problem according to the Bayes decision theory and to the related estimates of class conditional densities (likelihood) and class prior probabilities (priors) [1]. If noisy samples are uniformly distributed over classes, the estimations of class conditional densities results corrupted, while the estimations of prior probabilities are not affected from the presence of mislabeled patterns. On the contrary, if noisy samples are not uniformly distributed over classes, both the estimations of prior probabilities and of class conditional densities are biased from mislabeled patterns. Therefore, we expect that supervised algorithms, which (explicitly or implicitly) consider the prior probabilities for the classification of a generic input pattern (e.g., Bayesian classifier, *k*-Nearest Neighbor (*k*-NN) [1]-[3]) are more sensitive to unbalanced noisy samples distributions over classes than other algorithms that take into account only the class conditional densities (e.g., Maximum Likelihood [1], [2]).

In this chapter we address the above-mentioned problems by the following: 1) presenting a novel context-sensitive semisupervised SVM (CS^{4}VM) classification algorithm, which is robust to noisy training sets, and 2) analyzing the effect of noisy training patterns and of their distribution on the classification accuracy of widely used supervised and semisupervised classifiers.

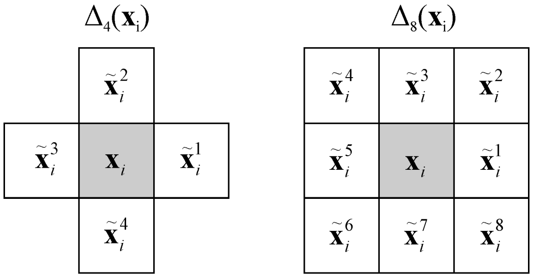
The choice of developing an SVM-based classifier is related to the important advantages that SVMs exhibit over other standard supervised algorithms [4]-[8]: 1) relatively high empirical accuracy and excellent generalization capabilities; 2) robustness to the Hughes phenomenon [9]; 3) convexity of the cost function used in the learning of the classifier; 4) sparsity of the solution; 5) possibility to use the kernel tricks for addressing non linear problems. In particular, the generalization capability of SVM (induced by the minimization of the structural risk) gives to SVM-based classifiers an intrinsic higher robustness to noisy training patterns than other standard algorithms that are based on the empirical risk minimization principle. In this framework, we propose an SVM-based technique for image classification especially developed to improve the robustness of standard SVM to the presence of noisy samples in the training set. The main idea behind the proposed CS^{4}VM is to exploit the spatial context information provided by the pixel belonging to the neighborhood system of each training sample (which are called context patterns) in order to contrast the bias effect due to the possible presence of mislabeled training patterns. This is achieved by both a semisupervised procedure (aiming to obtain the semilabels for context patterns) and the definition of a novel contextual term in the cost function associated with the learning of the CS^{4}VM. It is worth noting that this use of the contextual information is completely different from that of traditional context-sensitive classifiers (e.g., [10]-[16]), where contextual information is exploited for regularizing classification maps in the decision phase.

Another important contribution of this work is to present an extensive experimental analysis to investigate and compare the robustness to noisy training sets of the proposed CS^{4}VM and of other conventional classifiers. In greater detail, we considered the (Gaussian) Maximum likelihood (ML) classifier (which is based on a parametric estimation of the class conditional densities and does not consider the prior probabilities of the classes), the *k*-NN classifier (which is based on a distribution free local estimation of posterior probabilities that implicitly considers the class prior probabilities); the standard SVM classifier and the progressive semisupervised SVM (PS^{3}VM) [17]. The five considered classification algorithms were tested on two different data sets: 1) a very high resolution (VHR) multispectral image acquired by the Ikonos satellite and 2) a medium resolution multispectral image acquired by Landsat 5 Thematic Mapper. The experimental analysis was carried out, considering training sets including different amounts of noisy samples having different distributions over the considered classes.

The chapter is organized into six sections. Section 4.2 presents the proposed context-sensitive semisupervised SVM (CS^{4}VM) technique. Section 4.3 describes the design of the experiments carried out with different classifiers. Section 4.4 and 4.5 illustrate the experimental results obtained on the Ikonos and Landsat data sets, respectively. Finally, section 4.6, after discussion, draws the conclusion of the chapter.

4.2 Proposed context-sensitive semisupervised SVM (CS^{4}VM)

Let $\mathcal{I}$ denote a multispectral *d*-dimensional image of size $I \times J$ pixels. Let us assume that a training set $T = \{\mathcal{X}, \mathcal{Y}\}$ made up of N pairs $(\mathbf{x}_i, y_i)_{i=1}^{N}$ is available, where $\mathcal{X} = \{\mathbf{x}_i \mid \mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^{N} \subset \mathcal{I}$ is a subset of $\mathcal{I}$ and $\mathcal{Y} = \{y_i\}_{i=1}^{N}$ is the corresponding set of labels. For the sake of simplicity, since SVMs are binary classifiers, we first focus the attention on the two-class case (the general multiclass case will be addressed later). Accordingly, let us assume that $y_i \in \{+1; -1\}$ is the binary label of the pattern $\mathbf{x}_i$. We also assume that a restricted amount δ of training samples $\mathbf{x}_i$ may be associated with wrong labels y_i, i.e., labels that do not correspond to the actual class of the considered pixel. Let $\Delta_M(\mathbf{x})$ represent a local neighborhood system (whose shape and size depend on the specific investigated image and application) of the generic pixel $\mathbf{x}$, where M indicates the number of pixels considered in the neighborhood. Generally $\Delta_M(\mathbf{x})$ is a first or second order neighborhood system (see Fig. 4.1). Let $\tilde{\mathcal{X}} = \{\tilde{\mathbf{x}}_i^j \mid \tilde{\mathbf{x}}_i^j \in \Delta_M(\mathbf{x}_i), \forall \mathbf{x}_i \in \mathcal{X}, j = 1, \ldots, M\}$ be the set of (unlabeled) context patterns $\tilde{\mathbf{x}}_i^j$ made up of the pixels belonging to the neighborhood $\Delta_M(\mathbf{x}_i)$ of the generic training sample $\mathbf{x}_i$. It is worth noting that adjacent training pixels belong to both $\mathcal{X}$ and $\tilde{\mathcal{X}}$.



a) b)

Fig. 4.1 – Examples of neighborhood systems for the generic training pixel $\mathbf{x}_i$. a) First order system $\Delta_4(\mathbf{x}_i)$. b) Second order system $\Delta_8(\mathbf{x}_i)$.

The idea behind the proposed methodology is to exploit the information of the context patterns $\tilde{\mathcal{X}}$ to reduce the bias effect of the δ mislabeled training patterns on the definition of the discriminating hyperplane of the SVM classifier, thus decreasing the sensitivity of the learning algorithm to unreliable training samples. This is accomplished by explicitly including the samples belonging to the neighborhood system of each training pattern in the definition of the cost function used for the learning of the classifier. These samples are considered by exploiting the labels derived through a semisupervised classification process (for this reason they are called semilabeled samples) [18]-[20]. The semilabeled context patterns have the effect to mitigate the bias introduced by noisy patterns adjusting the position of the hyperplane. This strategy is defined according to a learning procedure for the proposed CS^{4}VM that is based on two main steps: 1) supervised learning with original training samples and classification of the (unlabeled) context patterns and 2) contextual semisupervised learning based on both original labeled patterns and semilabeled context patterns according to a novel cost function. These two steps are described in detail in the following subsections.

4.2.1 Step 1 - supervised learning and classification of context patterns

In the first step, a standard supervised SVM is trained by using the original training set T in order to classify the patterns belonging to the neighborhood system of each training pixels. The learning is performed according to the soft margin SVM algorithm, which results in the following constrained minimization problem:

$$\begin{cases} \min\limits_{\mathbf{w},b,\boldsymbol{\xi}} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \right\} \\ y_i \cdot \left[\mathbf{w}\cdot\Phi(\mathbf{x}_i) + b \right] \geq 1 - \xi_i \qquad \forall i = 1,\ldots,N \\ \xi_i \geq 0 \end{cases} \tag{4.1}$$

where $\mathbf{w}$ is a vector normal to the separation hyperplane, b is a constant such that $b/\|\mathbf{w}\|$ represents the distance of the hyperplane from the origin, $\Phi(\cdot)$ is a non-linear mapping function, ξ_i are slack variables that control the empirical risk (i.e., the number of training errors), and $C \in \mathbb{R}_0^+$ is a regularization parameter that tunes the tradeoff between the empirical error and the complexity term (i.e., the generalization capability). The above minimization problem can be rewritten in the dual formulation by using the Lagrange optimization theory, which leads to the following dual representation:

$$\begin{cases} \max\limits_{\boldsymbol{\alpha}} \left\{ \sum\limits_{i=1}^{N} \alpha_i - \frac{1}{2} \sum\limits_{i=1}^{N} \sum\limits_{j=1}^{N} y_i y_j \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \right\} & \\ \sum\limits_{i=1}^{N} y_i \alpha_i = 0 & \forall i = 1, \ldots, N \\ 0 \le \alpha_i \le C & \end{cases} \tag{4.2}$$

where α_i are the Lagrange multipliers associated with the original training patterns $\mathbf{x}_i \in \mathcal{X}$, and $k(\cdot,\cdot)$ is a kernel function such that $k(\cdot,\cdot) = \Phi(\cdot)\Phi(\cdot)$. The kernel function is used for implicitly mapping the input data into a high dimensional feature space without knowing the function $\Phi(\cdot)$ and still maintaining the convexity of the objective function [6]. Once α_i ($i = 1, ..., N$) are determined, each context pattern $\tilde{\mathbf{x}}_i^j$ in the neighborhood system $\Delta_M(\mathbf{x}_i)$ of the training pattern $\mathbf{x}_i$ is associated with a semilabel $\tilde{y}_i^j$ according to:

$$\hat{\tilde{y}}_i = \operatorname{sgn}\left[\sum_{n=1}^{N} y_n \alpha_n k\left(\mathbf{x}_n, \tilde{\mathbf{x}}_i^j\right) + b \right] \quad \forall \mathbf{x}_n \in \mathcal{X},\ \forall \tilde{\mathbf{x}}_i^j \in \tilde{\mathcal{X}} \tag{4.3}$$

where, given $f(\mathbf{x}) = \sum_{i=1}^{N} y_i \alpha_i k\left(\mathbf{x}_i, \mathbf{x}\right) + b$, b is chosen so that $y_i f(\mathbf{x}_i) = 1$ for any i with $0 < \alpha_i < C$.

4.2.2 Step 2 - context-sensitive semisupervised learning

Taking into account the semilabels (i.e., the labels obtained in the previous step) of the context patterns belonging to $\tilde{\mathcal{X}}$, we define the following novel context-sensitive cost function for the learning of the classifier:

$$\Psi(\mathbf{w}, \boldsymbol{\xi}, \psi) = \frac{1}{2} \left\| \mathbf{w} \right\|^2 + C \sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} \sum_{j=1}^{M} \kappa_i^j \psi_i^j \tag{4.4}$$

where ψ_i^j are context slack variables and $\kappa_i^j \in \mathbb{R}_0^+$ are parameters that permit to weight the importance of context patterns (see Fig. 4.2). The resulting constrained minimization problem associated with the learning of the CS^{4}VM is the following:

$$\begin{cases} \min\limits_{\mathbf{w}, b, \boldsymbol{\xi}, \psi} \Psi(\mathbf{w}, \boldsymbol{\xi}, \psi) & \\ y_i \cdot \left[\mathbf{w} \cdot \Phi\left(\mathbf{x}_i\right) + b \right] \ge 1 - \xi_i & \\ \tilde{y}_i^j \cdot \left[\mathbf{w} \cdot \Phi\left(\tilde{\mathbf{x}}_i^j\right) + b \right] \ge 1 - \psi_i^j & \forall i = 1, \ldots, N \\ \psi_i^j, \xi_i \ge 0 & \forall j = 1, \ldots, M \end{cases} \tag{4.5}$$

The cost function in (4.4) contains a novel contextual term (made up of $N \cdot M$ elements) whose aim is to regularize the learning process with respect to the behavior of the context patterns in the neighborhood of the training pattern under consideration. The rationale of this term is to balance the contribution of possibly mislabeled training samples according to the semilabeled pixels of the neighborhood. The context slack variables $\psi_i^j = \psi_i^j\left(\tilde{\mathbf{x}}_i^j, \tilde{y}_i^j, \mathbf{w}, b\right)$ depend on

$\tilde{\mathbf{x}}_i^j \in \Delta_M(\mathbf{x}_i)$ and, accordingly, permit to directly take into account the contextual information in the learning phase. They are defined as:

$$\psi_i^j = \max\left\{0, 1 - \tilde{y}_i^j \cdot \left[\mathbf{w} \cdot \Phi\left(\tilde{\mathbf{x}}_i^j\right) + b\right]\right\} \quad \forall i = 1, \ldots, N, \forall j = 1, \ldots, M \tag{4.6}$$

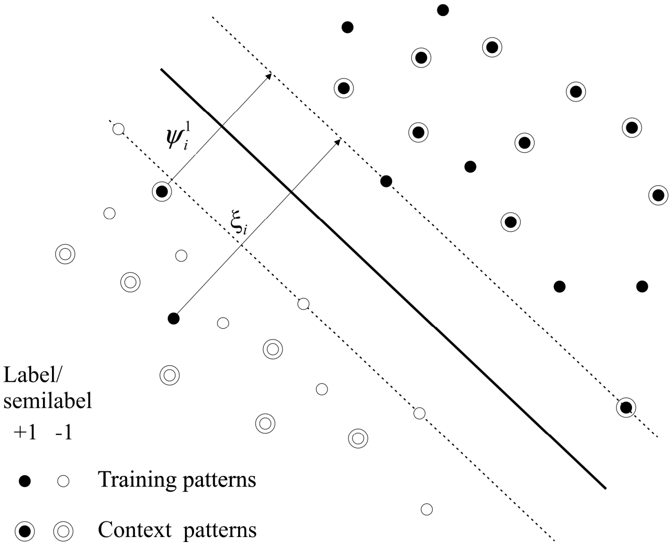


Fig. 4.2 – Example of training and related context patterns in the kernel-induced feature space.

The parameters $\kappa_i^j \in \mathbb{R}_0^+$ weight the context patterns $\tilde{\mathbf{x}}_i^j$ depending on the agreement of their semilabels $\tilde{y}_i^j$ with that of the related training sample y_i. The hypothesis at the basis of the weighting system of the context patterns is that the pixels in the same neighborhood system have high probability to be associated to the same information class (i.e., the labels of the pixels are characterized by high spatial correlation). In particular, κ_i^j are defined as follows:

$$\kappa_i^j = \begin{cases} \kappa_1 & \text{if} \quad y_i = \tilde{y}_i^j \\ \kappa_2 & \text{if} \quad y_i \neq \tilde{y}_i^j \end{cases} \tag{4.7}$$

where κ_1 and κ_2 are chosen from the user. The role of κ_1 and κ_2 is to define the importance of the context patterns. In particular, it is very important to define the ratios C/κ_i, $i = 1, 2$ which tune the weight of context patterns with respect to the patterns of the original training set. According to our hypothesis, in order to adequately penalize the mislabeled training patterns, it is suggested to fix $\kappa_1 \geq \kappa_2$ as, in general, contextual patterns whose semilabels are in agreement with the label of the related training pattern should be considered more reliable than those whose semilabels are different. The selection of κ_1 and κ_2 can be simplified fixing *a priori* the ratio $\kappa_1 / \kappa_2 = \mathrm{K}$, thus focusing the attention only on κ_1 or on the ratio C/κ_1.

It is worth noting that the novel cost function defined in (4.4) maintains the important property of convexity of the cost function of the standard SVM. This allows us to solve the problem according to quadratic programming algorithms. By properly adjusting the Karush-Kuhn-Tucker conditions [i.e., the necessary and sufficient conditions for solving (4.5)], we derived the following dual bound maximization problem:

$$\begin{cases} \max\limits_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N}\left(\alpha_i + \sum_{j=1}^{M}\beta_i^j\right) - \frac{1}{2}\sum_{i=1}^{N}\sum_{h=1}^{N}\begin{bmatrix} y_i y_h \alpha_i \alpha_h k\left(\mathbf{x}_i, \mathbf{x}_h\right) + \\ +2 y_i \alpha_i \sum_{j=1}^{M} \tilde{y}_h^i \beta_h^i k\left(\mathbf{x}_i, \tilde{\mathbf{x}}_h^j\right) + \\ +\sum_{q=1}^{M}\sum_{j=1}^{M} \tilde{y}_i^q \tilde{y}_h^j \beta_i^q \beta_h^j k\left(\tilde{\mathbf{x}}_i^q, \tilde{\mathbf{x}}_h^j\right) \end{bmatrix} \right\} \\ \sum_{i=1}^{N}\left(y_i \alpha_i + \sum_{j=1}^{M}\tilde{y}_i^j \beta_i^j\right) = 0 \\ 0 \le \alpha_i \le C \qquad \forall i = 1, \dots, N \\ 0 \le \beta_i^j \le \kappa_i^j \qquad \forall j = 1, \dots, M \end{cases} \tag{4.8}$$

where α_i and r_i are the Lagrange multipliers associated with original training patterns, while β_i^j and s_i^j are the Lagrange multipliers associated with contextual patterns. The Lagrange multipliers α_i associated with the original labeled patterns are superiorly bounded by C (they all have the same importance). The upper bound for the Lagrange multipliers β_i^j associated with context patterns is κ_i^j, as it comes from (4.7). Once determined α_i and β_i^j ($i = 1,..., N$, $j = 1,..., M$) the generic pattern **x** belonging to the investigated image $\mathcal{I}$ can be classified according to the following decision function:

$$\hat{y} = \operatorname{sgn}\left\{ \sum_{i=1}^{N}\left[y_i \alpha_i k\left(\mathbf{x}_i, \mathbf{x}\right) + \sum_{j=1}^{M} \tilde{y}_i^j \beta_i^j k\left(\tilde{\mathbf{x}}_i^j, \mathbf{x}\right) \right] + b \right\} \quad \forall \mathbf{x}_i \in \mathcal{X}, \quad \forall \tilde{\mathbf{x}}_i^j \in \tilde{\mathcal{X}} \tag{4.9}$$

where, given $f(\mathbf{x}) = \sum_{i=1}^{N}\left[y_i \alpha_i k\left(\mathbf{x}_i, \mathbf{x}\right) + \sum_{j=1}^{M} \tilde{y}_i^j \beta_i^j k\left(\tilde{\mathbf{x}}_i^j, \mathbf{x}\right) \right] + b$, b is chosen so that $y_i f(\mathbf{x}_i) = 1$ for any *i* with $0 < \alpha_i < C$, and $\tilde{y}_i^j f(\tilde{\mathbf{x}}_i^j) = 1$ for any *i* and *j* with $0 < \beta_i^j < \kappa_i^j$.

It is worth noting that the proposed formulation could be empirically defined by considering different analytical forms for the kernels associated with the original training samples and the context patterns (composite kernel approach). From a general perspective, this would increase the flexibility of the method. However, as the training patterns and the context patterns are represented by the same feature vectors, the use of composite kernels (which would result in a further increase of the number of free parameters to set in the leaning of the classifier, and thus, in an increase of the computational cost required from the model-selection phase) does not seem useful.

4.2.3 Multiclass architecture

Let us extend the binary CS^{4}VM to the solution of multiclass problems. Let $\Omega = \{\omega_1, ..., \omega_L\}$ be the set of *L* information classes that characterize the considered problem. As for the conventional SVM, the multiclass problem should be addressed with a structured architecture made up of binary classifiers. However, the properties of CS^{4}VM lead to an important difference with respect to the standard supervised SVM. This difference is related to the *step 2* of the learning of the CS^{4}VM. In this step we assume to be able to give a reliable label to all patterns in the neighborhood system of each training pattern. In order to satisfy this constraint, we should define binary classification problems for each CS^{4}VM included in the multiclass architecture characterized from an exhaustive representation of classes.

Let each CS^{4}VM of the multiclass architecture solve a binary subproblem, where each pattern should belong to one of the two classes Ω_A or Ω_B, defined as proper subsets of the original set of labels Ω. The contextual semisupervised approach requires that, for each binary CS^{4}VM of the multiclass architecture, there must be an exhaustive representation of all possible labels, i.e.,

$$\Omega_A \cup \Omega_B = \Omega \tag{4.10}$$

If (4.10) is not satisfied, some semilabels of context patterns $\tilde{\mathbf{x}}_i^j$ may not be represented in the binary sub-problem and the context sensitive semisupervised learning can not be performed. According to this constraint, we propose to adopt a one-against-all (OAA) multiclass architecture, which is made up of *L* parallel CS^{4}VM, as shown in Fig. 4.3.

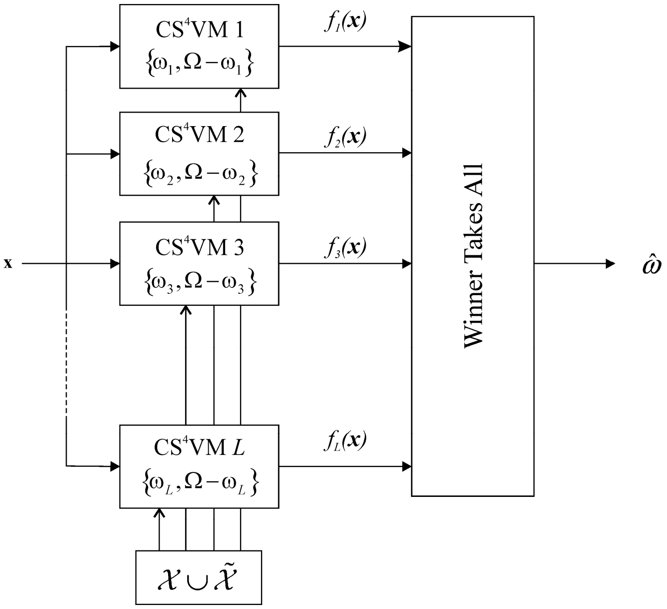

CS^{4}VM 1
$\{\omega_1, \Omega-\omega_1\}$
$f_1(\boldsymbol{x})$
CS^{4}VM 2
$\{\omega_2, \Omega-\omega_2\}$
$f_2(\boldsymbol{x})$
x
CS^{4}VM 3
$\{\omega_3, \Omega-\omega_3\}$
$f_3(\boldsymbol{x})$
Winner Takes All
$\hat{\omega}$
CS^{4}VM L
$\{\omega_L, \Omega-\omega_L\}$
$f_L(\boldsymbol{x})$
$\mathcal{X} \cup \tilde{\mathcal{X}}$

Fig. 4.3 – OAA architecture for addressing the multiclass problem with the proposed CS^{4}VM.

The *i*-th CS^{4}VM solves a binary problem defined by the information class $\{\omega_i\} \in \Omega$ against all the others $\Omega - \{\omega_i\}$. In this manner all the binary sub-problems of multiclass architecture satisfy (4.10). The “winner-takes-all” rule is used for taking the final decision, i.e.,

$$\hat{\omega} = \underset{i=1,\dots,L}{\arg\max} \{ f_i(\mathbf{x}) \} \tag{4.11}$$

where $f_i(\mathbf{x})$ represent the output of the *i*-th CS^{4}VM.

It is worth noting that other multiclass strategies that are commonly adopted with standard SVM [such as the one-against-one (OAO)] [21], cannot be used with the proposed CS^{4}VM as do not satisfy (4.10). Nevertheless, other multi-class architectures could be specifically developed for the CS^{4}VM approach, which should satisfy the constraint defined in (4.10).

4.3 Design of experiments

In this section, we describe the extensive experimental phase carried out to evaluate the robustness to the presence of noisy training samples of the proposed CS^{4}VM and of other standard supervised and semisupervised classification algorithms. In particular, we compare the accuracy (in terms of kappa coefficient [22]) obtained by the proposed CS^{4}VM with those yielded by other classification algorithms: the progressive semisupervised SVM (PS^{3}VM) [17], the standard supervised SVM, the Maximum Likelihood (ML), and the *k*-Nearest Neighbors (*k*-NN). We carried out different kinds of experiments by training the classifiers: 1) with the original training samples (with their correct labels), and 2) with different synthetic training sets, where mislabeled patterns (i.e., patters with wrong labels) were added to the original training set in different percentages (10%, 16%, 22%, 28%) with respect to the total number of training samples. In the second kind of experiments, we manually introduced mislabeled training samples considering the particular scene under investigation and simulating realistic mislabeling errors (e.g., caused by possible photointerpretation errors). The spatial location of wrong samples was distributed over the whole scene, by considering also clusters of pixels in the same neighborhood system. We analyzed the effects of noisy training sets on the classification accuracy, in two different scenarios (which simulate different kinds of mislabeling errors): a) wrong samples are uniformly added to all the information classes (thus simulating the presence of mislabeling errors in the training points that does not depend on the land cover type); b) wrong patterns are added to one specific class or to a subset of the considered classes (thus simulating a systematic error in the collection of ground truth samples for specific land cover types).

In all the experiments, for the ML classifier we adopted the Gaussian function as model for the probability density functions of the classes. Concerning the *k*-NN classification algorithm, we carried out several trials, varying the value of *k* from 1 to 40 in order to identify the value that maximizes the kappa accuracy on the test set.

For the SVM-based classifiers (CS^{4}VM, PS^{3}VM and standard SVM) we employed the Sequential Minimal Optimization (SMO) algorithm [23] (with proper modifications for the CS^{4}VM) and used Gaussian kernel functions (ruled by the free parameter σ that expresses the width of the Gaussian function). All the data were normalized to a range [0, 1] and the model selection for deriving the learning parameters was carried out according to a grid-search strategy on the basis of the kappa coefficient of accuracy obtained on the test set.

For the standard SVM, the value of $2\sigma^2$ was varied in the range [10^{-2}, 10], while the values of *C* were concentrated in the range [20, 200] after a first exploration in a wider range. For the model selection of both the CS^{4}VM and the PS^{3}VM, we considered the same values for C and $2\sigma^2$ as for the SVM in order to have comparable results. Moreover, for the proposed CS^{4}VM we fixed the value of $\mathrm{K} = \kappa_1 / \kappa_2 = 2$ and used the following values for C/κ_1: 2, 4, 6, 8, 10, 12, 14. For the definition of the context patterns we considered a first order neighborhood system. With regard to the PS^{3}VM, the value of $C^{*(0)}$ was varied in the range [0.1,1], the one of γ was varied in the range [10,100], and ρ was varied in the range [10, 100].

For simplicity, the model selection for all the SVM-based classifiers and the *k*-NN algorithm was carried out on the basis of the kappa coefficient of accuracy computed on the test set, which does not contain mislabeled samples. It is worth noting that this does not affect the relative results of the comparison, as the same approach was used for all the classifiers. It is important to observe that the proposed CS^{4}VM method does not rely on the assumption of noise-free samples in the test set for parameter settings. The use of context patterns is effective in mitigating the bias effect introduced by noisy patterns even if the selected model is optimized on a noisy test set. In this condition, we may have an absolute decrease of classification accuracy, but the capability to mitigate the effects of wrong samples on the final classification result does not change.

In the experiments, we considered two data sets: the first one is made up of a very high geometrical resolution multispectral image acquired by the Ikonos satellite over the city of Ypenburg (The Netherlands); the second one is made up of a medium resolution multispectral image acquired by the sensor Thematic Mapper of Landsat 5 in the surroundings of the city of Trento (Italy). The results obtained on the two data sets are presented in the following two sections.

4.4 Experimental results: Ikonos data set

The first considered data set is made up of the first three bands (corresponding to visible wavelengths) of an Ikonos sub-scene of size 387 × 419 pixels (see Fig. 4.4). The 4 m spatial resolution spectral bands have been reported to a 1 m spatial resolution according to the Gram-Schmidt pansharpening procedure [24]. The available ground truth (which included the information classes grass, road, shadow, small-aligned building, white-roof building, gray-roof building and red-roof building) collected on two spatially disjoint areas was used to derive a training set and a test set for the considered image (see Table 4.1). This setup allowed us to study the generalization capability of the systems by performing validation on areas spatially disjoint from those used in the learning of the classification algorithm. This is very important because of the nonstationary behavior of the spectral signatures of classes in the spatial domain. Starting from the original training set, several data sets were created adding different percentages of mislabeled pixels in order to simulate noisy training sets as described in the previous section.

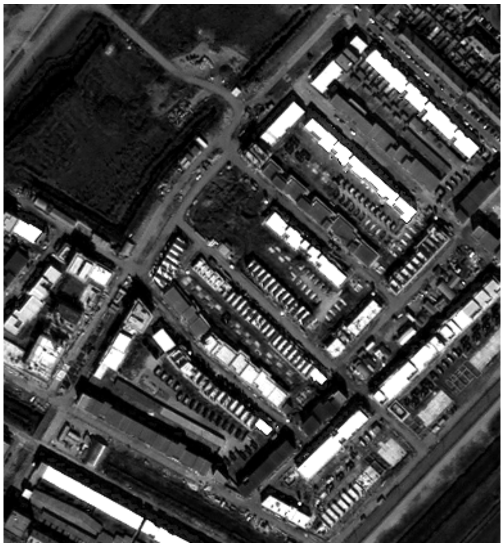


Fig. 4.4 - Band 3 of the Ikonos image.

Table 4.1- Number of patterns in the training and test sets (Ikonos data set).

Class		**Number of patterns**	
		Training Set	**Test Set**
Grass		63	537
Road		82	376
Building	Small-aligned	62	200
	White-roof	87	410
	Gray-roof	65	336
	Red-roof	19	92
Shadow		30	231

4.4.1 Results with mislabeled training patterns uniformly added to all classes

In the first set of experiments, different percentages (10%, 16%, 22%, 28%) of mislabeled patterns (with respect to the total number of samples) were uniformly added to all classes of the training set. The accuracy yielded on the test set by all the considered classifiers versus the percentage of mislabeled patterns are reported in Table 4.2 and plotted in Fig. 4.5. As one can see, with the original training set, the proposed CS^{4}VM exhibited higher kappa coefficient of accuracy than the other classifiers. In greater detail, the kappa coefficient obtained with the CS^{4}VM is slightly higher than the ones obtained with the standard SVM and the PS^{3}VM (+1.6%), and sharply higher than those yielded by the *k*-NN (+6.6%) and the ML (+8%). This confirms that the semisupervised exploitation of contextual information of training patterns allows us increasing the classification accuracy (also if their labels are correct). In this condition, the PS^{3}VM classifier did not increase the classification accuracy of the standard SVM. When mislabeled samples were added to the original training set, the accuracies obtained with ML and *k*-NN classifiers sharply decreased, whereas SVM-based classifiers showed to be much more robust to "noise" (by increasing the number of mislabeled samples the kappa accuracy decreased slowly). In greater detail, the kappa accuracy of the ML classifier decreased of 15.9% in the case of 10% of mislabeled samples with respect to the result obtained in the noise-free case, while the *k*-NN reduced its accuracy by 5.8% in the same condition. More generally, the *k*-NN classifier exhibited higher and more stable accuracies than the ML with all the considered amounts of noisy patterns. In all the considered trials, the proposed CS^{4}VM exhibited higher accuracy than the other classifiers. In addition, with moderate and large numbers of mislabeled patterns (16%, 22% and 28%), it was more stable than the SVM and the PS^{3}VM. In the trials with noisy training sets the PS^{3}VM classifier slightly increased the accuracy obtained by the standard SVM.

Table 4.2- Kappa coefficient of accuracy on the test set with different percentages of mislabeled patterns added uniformly to the training set (Ikonos data set).

% of mislabeled patterns	**Kappa Accuracy**				
	CS^{4}VM	**PS^{3}VM**	**SVM**	***k*-NN**	**ML**
0	0.927	0.907	0.907	0.861	0.847
10	0.919	0.910	0.907	0.803	0.688
16	0.921	0.869	0.866	0.787	0.801
22	0.893	0.862	0.861	0.781	0.727
28	0.905	0.874	0.860	0.763	0.675

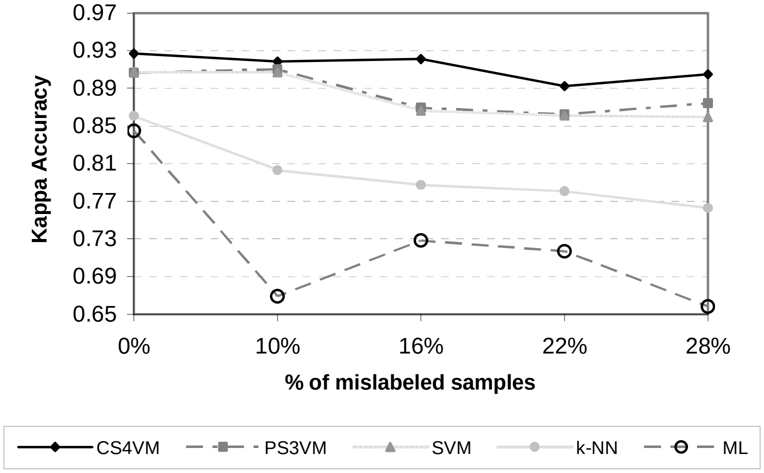


Fig. 4.5 – Behavior of the kappa coefficient of accuracy on the test set versus the percentage of mislabeled training patterns uniformly distributed over all classes introduced in the training set (Ikonos data set).

In order to better analyze the results of SVM and CS^{4}VM, we compared the average and the minimum kappa accuracies of the binary classifiers that made up the OAA multi-class architecture (see Fig. 4.6 and Table 4.3). It is possible to observe that the average kappa accuracy of the binary CS^{4}VMs was higher than that of the binary SVMs, and exhibited a more stable behavior when the amount of noise increased. Moreover, the accuracy of the class most affected by the inclusion of mislabeled patterns in the training set was very stable with the proposed classification algorithm, whereas it sharply decreased with the standard SVM when large percentages of mislabeled patterns were included in the training set. This confirms the effectiveness of the proposed CS^{4}VM, which exploits the contributions of the contextual term (and thus of contextual patterns) for mitigating the effects introduced by the noisy samples.

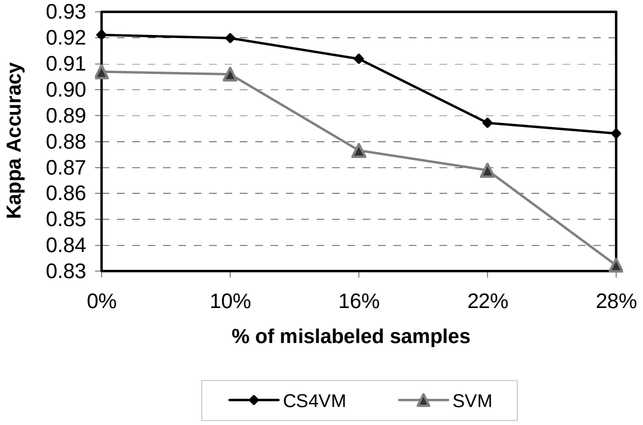


Fig. 4.6 - Behavior of the average kappa coefficient of accuracy (computed on all the binary CS^{4}VMs and SVMs included in the multiclass architecture) versus the percentage of mislabeled training patterns uniformly added to all classes (Ikonos data set).

Table 4.3 - Kappa coefficient of accuracy exhibited from the binary CS^{4}VM and SVM that resulted in the lowest accuracy among all binary classifiers included in the multiclass architecture versus the percentages of mislabeled training patterns uniformly added to all classes (Ikonos data set).

% of mislabeled patterns	**Kappa Accuracy**		
	CS^{4}VM	**SVM**	**Δ(%)**
0	0.783	0.756	2.7
10	0.784	0.767	1.8
16	0.757	0.738	1.9
22	0.751	0.691	6.0
28	0.755	0.509	24.6

4.4.2 Results with mislabeled training patterns concentrated on specific classes

In the second set of experiments, several samples of the class “grass” were added to the original training set with the wrong label “road” in order to reach 10% and 16% of noisy patterns. In addition “white-roof building” patterns were included with label “grey-roof building” to reach 22% and 28% of noisy samples. The resulting classification problem proved quite critical, as confirmed by the significant decrease in the kappa accuracies yielded by the considered classification algorithms (see Fig. 4.7 and Table 4.4). Nevertheless, also in this case, the context-based training of the CS^{4}VM resulted in a significant increase of accuracy with respect to other classifiers. The kappa accuracy of the *k*-NN classifier dramatically decreased when the percentage of noisy patterns increased (in the specific case of 28% of mislabeled samples the kappa accuracy decreased of 35.1% with respect to the original training set). The ML decreased its accuracy of 10.1% with 10% of noisy patterns, but exhibited a more stable behavior with respect to the *k*-NN when the amount of noisy patterns was further increased. The standard SVM algorithm obtained accuracies higher than those yielded by the *k*-NN and ML classifiers, while the PS^{3}VM classifier in general slightly improved the accuracy of the standard SVM. However, with 28% of noisy patterns, the kappa accuracy sharply decreased to 0.629 (below the performance of ML). This behavior was strongly mitigated by the proposed CS^{4}VM (which exhibited a kappa accuracy of 0.820 in the same conditions).

Table 4.4 - Kappa coefficient of accuracy on the test set with different percentages of mislabeled patterns added to specific classes of the training set (Ikonos data set).

% of mislabeled patterns	**Kappa Accuracy**				
	CS^{4}VM	**PS^{3}VM**	**SVM**	***k*-NN**	**ML**
0	0.927	0.907	0.907	0.861	0.847
10	0.906	0.855	0.841	0.690	0.746
16	0.781	0.769	0.765	0.672	0.734
22	0.828	0.767	0.762	0.525	0.722
28	0.820	0.632	0.629	0.510	0.721

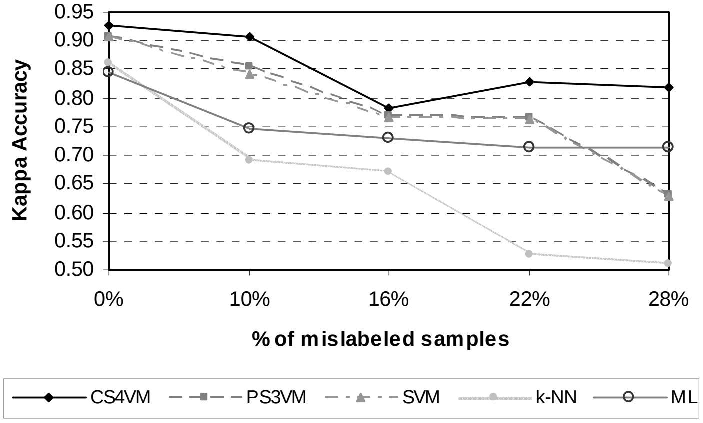


Fig. 4.7 – Behavior of the kappa coefficient of accuracy on test set versus the percentage of mislabeled training patterns concentrated on specific classes of the training set (Ikonos data set).

Considering the behavior of the average kappa of the binary SVMs and CS^{4}VMs that made up the OAA multi-class architecture (see Fig. 4.8), it is possible to note that the CS^{4}VM always improved the accuracy of the standard SVM, and the gap between the two classifiers increased by increasing the amount of noisy samples. In the very critical case of 28% of mislabeled patterns, the context-based learning of CS^{4}VM improved the average kappa accuracy of binary SVMs by 9.2%. Moreover, the kappa coefficient of the class with the lowest accuracy with the proposed CS^{4}VM, even if small, was sharply higher than that of the standard SVM in all the considered trials (see Table 4.5). This behavior shows that on this data set the proposed method always improved the accuracy of the most critical binary classifier.

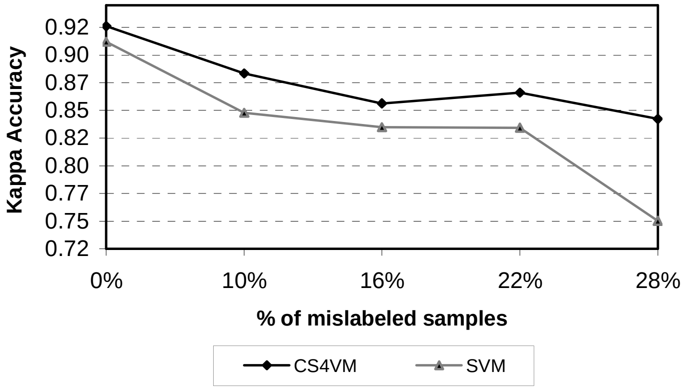


Fig. 4.8 - Behavior of the average kappa coefficient of accuracy (computed on all the binary CS^{4}VMs and SVMs included in the multi-class architecture) versus the percentage of mislabeled training patterns concentrated on specific classes (Ikonos data set).

Table 4.5- Kappa coefficient of accuracy exhibited from the binary CS^{4}VM and SVM that resulted in the lowest accuracy among all binary classifiers included in the multiclass architecture versus the percentages of mislabeled training patterns concentrated on specific classes (Ikonos data set).

% of mislabeled patterns	**Kappa Accuracy**		
	CS^{4}VM	**SVM**	**Δ(%)**
0	0.783	0.756	2.7
10	0.620	0.422	19.8
16	0.449	0.360	8.9
22	0.538	0.360	17.8
28	0.450	0.360	9.0

Fig. 4.9 shows the classification maps obtained training the considered classifiers with 28% of mislabeled patterns added on specific classes (“roads” and “grey roof buildings”) of the training set (the map obtained with the PS^{3}VM is not reported because it is very similar to the one yielded with the SVM classifier). As one can see, in the classification maps obtained with the SVM, the k-NN, and the ML algorithms, many pixels of the class grass are confused with the class road, while white roof buildings are confused with grey roof buildings. This effect is induced by the presence of noisy training samples affecting the aforementioned classes. In grater detail, the SVM classifier was unable to correctly recognize the red roof buildings, while the k-NN technique often misrecognized the shadows present in the scene as red roof buildings and white roof buildings as grey roof buildings. Moreover the thematic map obtained with the k-NN is very noisy and fragmented (as confirmed by the low kappa coefficient of accuracy). The thematic map obtained with the proposed CS^{4}VM clearly appears more accurate and less affected by the presence of mislabeled patterns.

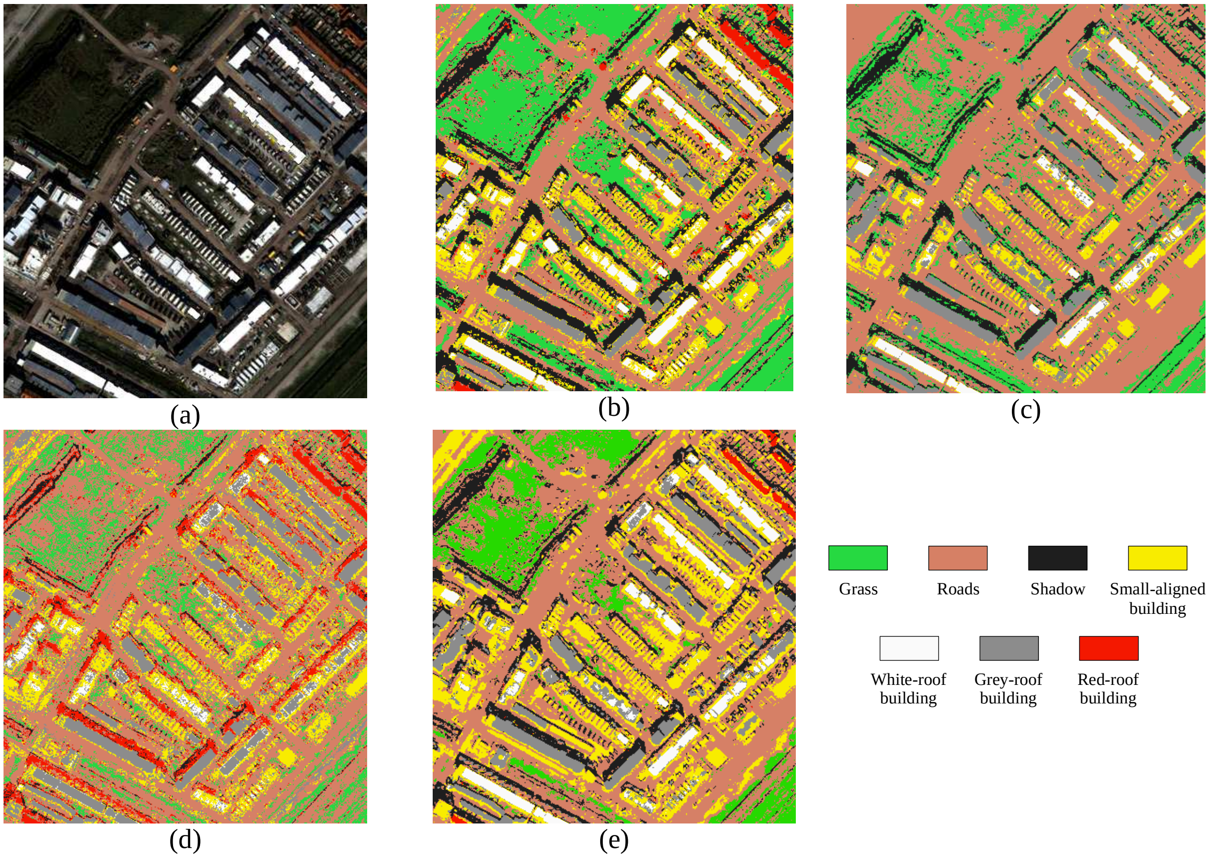

(a) (b) (c) (d) (e)

Fig. 4.9 – (a) True color composition of the Ikonos image. Classification maps obtained by the different classifiers with the training set containing 28% of mislabeled patterns added on specific classes. (b) CS^{4}VM. (c) SVM. (d) *k*-NN. (e) ML.

4.5 Experimental results: Landsat data set

The second data set consists of an image acquired by the Landsat 5 TM sensor with a GIFOV of 30 m. The considered image has size of 1110 × 874 pixels and was taken in the surrounding of the city of Trento (Italy) (see Fig. 4.10). A six-class classification problem (with forest, water, urban, rock, fields, and grass classes) was defined according to the available ground truth collected on two spatially disjoint areas and used to derive the training and test sets (see Table 4.6). As for the Ikonos data set, this setup allowed us to study the generalization capability of the algorithms by classifying areas spatially disjoint from those used in the learning of the classifier. The important difference between this data set and the previous one consists in the geometric resolution, which in this case is significantly smaller than in the previous case (30 m vs. 1 m). Similarly to the Ikonos data set, several noisy training sets were created adding different amount of mislabeled pixels to the original data set: 1) with uniform distribution over the classes and 2) concentrated on a specific class.

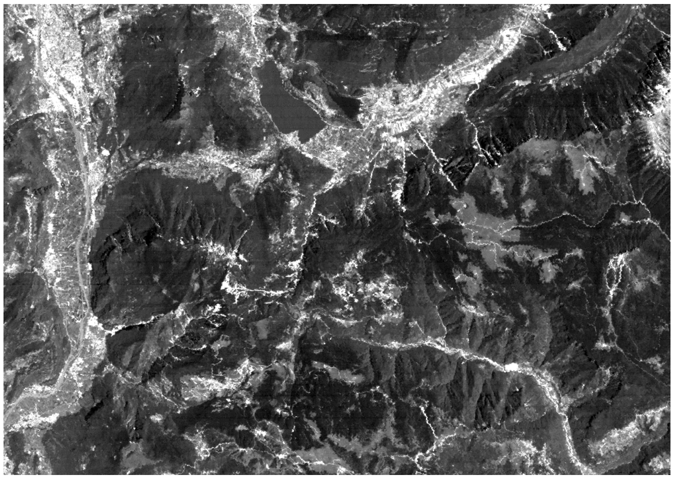


Fig. 4.10 - Band 2 of the Landsat TM multispectral image.

Table 4.6- Number of patterns in the training and test set (Landsat data set).

Class	Number of patterns	
	Training Set	**Test Set**
Forest	128	538
Water	118	177
Urban	137	289
Rocks	45	51
Fields	93	140
Grass	99	227

4.5.1 Results with mislabeled training patterns uniformly added to all classes

Table 4.7 shows the accuracies obtained in the first set of experiments where mislabeled patterns were uniformly added to the information classes. Fig. 4.11 depicts the behavior of the kappa accuracy versus the number of mislabeled patterns included in the training set for all the considered classifiers. It is possible to observe that with the noise-free training set, the proposed CS^{4}VM led to the highest accuracy, slightly improving the kappa coefficient of standard SVM by 0.8%. The ML classifier performed very well with the noise-free training set (the kappa accuracy was 0.923), but decreased its accuracy to 0.778 when only 10% of mislabeled patterns were introduced in the original training set, and its accuracy further decreased to 0.691 when the mislabeled samples reached 16%. The *k*-NN classifier led to lower accuracy than the ML in absence of noise, but showed to be less sensitive to noisy patterns uniformly added to the training set, thus exhibiting a more stable behavior. On the contrary, SVM-based classification algorithms proved to be robust to the presence of mislabeled training samples. Indeed, the excellent generalization capability of the SVM led to even slightly increase the classification accuracy when a small amount of mislabeled patterns was added to the training set. The PS^{3}VM algorithm resulted in a small improvement with respect to the SVM classifier in the trials where mislabeled samples were added to the training set. The kappa accuracy of the SVM classifier slightly decreased when the mislabeled samples exceeded 16%, reducing its accuracy by 3% with respect to the noise-free case. In these cases the proposed CS^{4}VM further enhanced the robustness of SVM, leading to kappa accuracies that were always above 0.91.

Table 4.7- Kappa coefficient of accuracy on test set using different percentages of mislabeled patterns added uniformly to the training set (Landsat data set).

% of mislabeled patterns	**Kappa Accuracy**				
	CS^{4}VM	**PS^{3}VM**	**SVM**	***k*-NN**	**ML**
0	0.927	0.915	0.919	0.912	0.923
10	0.930	0.935	0.931	0.905	0.778
16	0.935	0.932	0.930	0.893	0.691
22	0.921	0.891	0.886	0.868	0.686
28	0.916	0.886	0.886	0.840	0.681

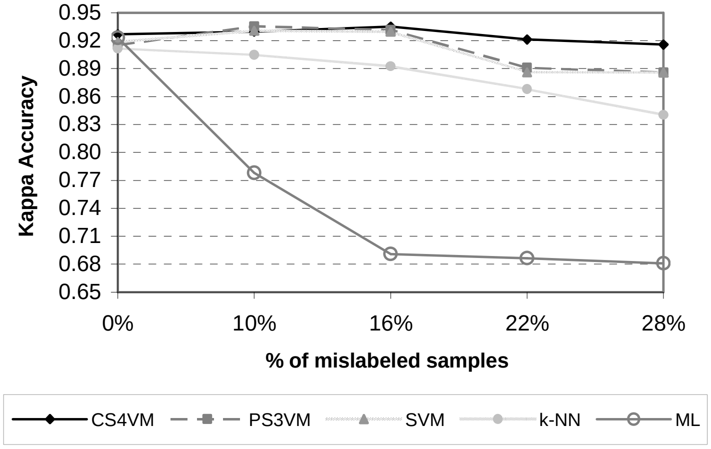


Fig. 4.11 – Behavior of the kappa coefficient of accuracy on test set versus the percentage of mislabeled training patterns uniformly added to all classes (Landsat data set).

This behavior is confirmed by the analysis of the average and minimum kappa computed on the binary classifiers (see Fig. 4.12 and Table 4.8), which highlights that the CS^{4}VM significantly improved the accuracy with respect to the SVM. Such an improvement was more significant when increasing the amount of noise; thus, the CS^{4}VM resulted in a more stable value of the kappa coefficient with respect to the percentage of mislabeled patterns present in the training set. It is worth noting that on this data set the proposed CS^{4}VM always improved the average kappa accuracy of the binary classifiers, even in cases where the global multiclass kappa coefficient of the CS^{4}VM was slightly smaller than the one obtained with the standard SVM. This can be explained observing that the decision strategy associated with the OAA multiclass architecture in some cases could "recover" the errors of binary classifiers by assigning the correct label to a pattern when comparing the output of binary classifiers. Nevertheless, the increased average accuracy of the binary CS^{4}VMs is an important property because involves more stable and reliable classification results.

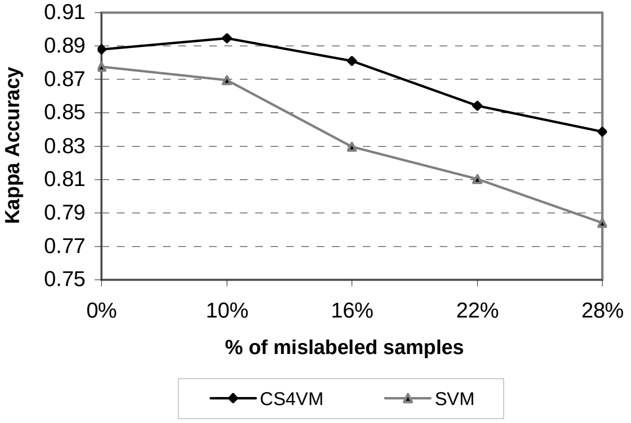


Fig. 4.12 - Behavior of the average kappa coefficient of accuracy (computed on all the binary CS^{4}VMs and SVMs included in the multiclass architecture) versus the percentage of mislabeled training patterns uniformly added to all classes (Landsat data set).

Table 4.8– Kappa coefficient of accuracy exhibited from the CS^{4}VM and SVM that resulted in the lowest accuracy among all binary classifiers included in the multiclass architecture versus the percentages of mislabeled training patterns uniformly added to all classes (Landsat data set).

% of mislabeled patterns	**Kappa Accuracy**		
	CS^{4}VM	**SVM**	**Δ(%)**
0	0.701	0.701	0.0
10	0.701	0.701	0.0
16	0.650	0.627	2.3
22	0.650	0.579	7.1
28	0.641	0.498	14.3

Fig. 4.13 shows the classification maps obtained training the classifiers with 28% of mislabeled patterns uniformly added to all the classes. It is possible to observe that the map generated by the proposed CS^{4}VM is the most accurate. In the maps yielded by the SVM, the *k*-NN, and the ML algorithms several pixels are misclassified as water (the map obtained with the PS^{3}VM is not reported as very similar to the SVM map). In grater detail, the map obtained with the *k*-NN presents confusion between the classes water and urban, and the classes forest and water. In the map obtained by the ML, grass areas are often confused with forest.

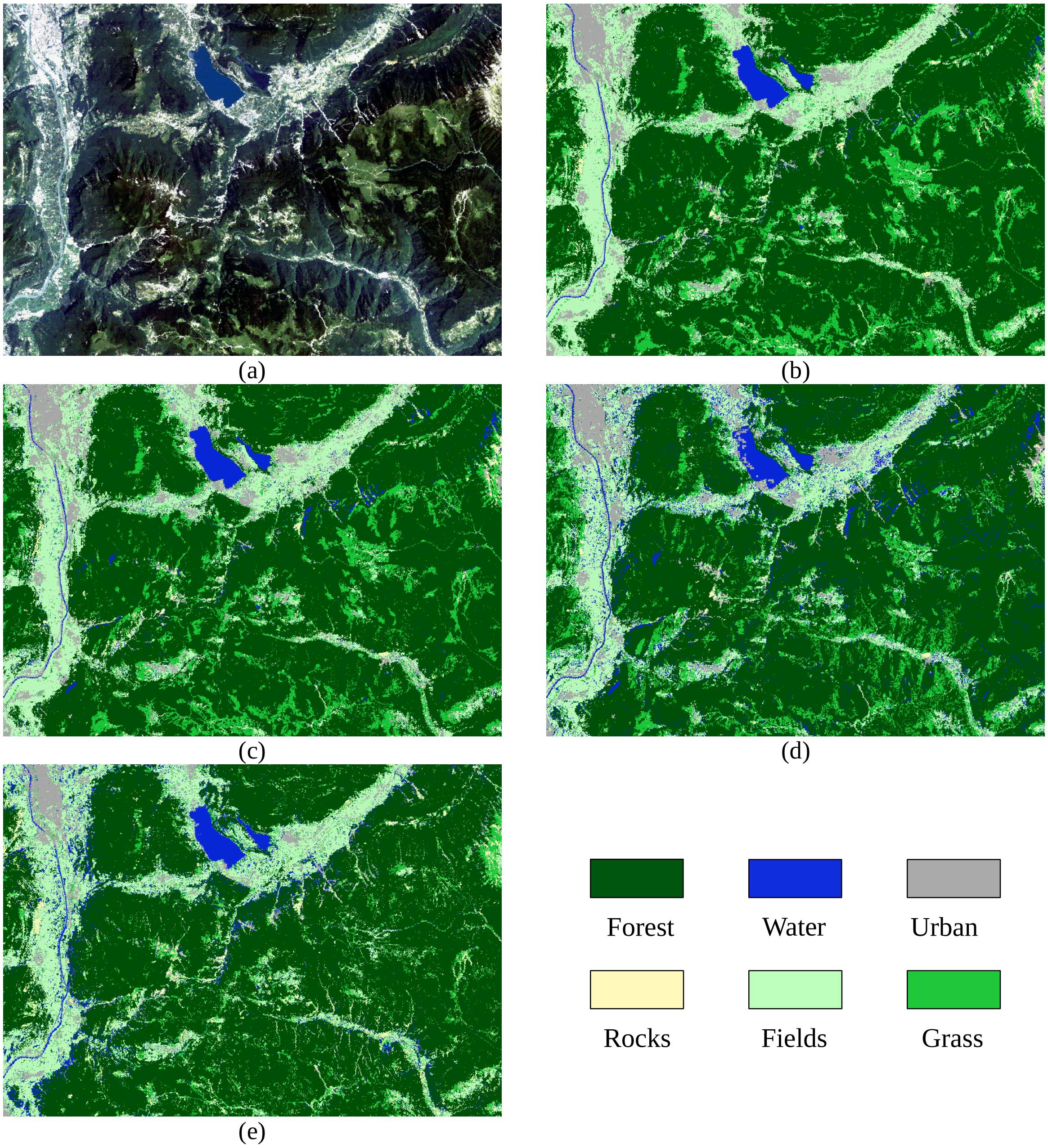

(a) (b) (c) (d) (e)

Fig. 4.13 – (a) True color composition of Landsat image. Classification maps obtained by the different classifiers with the training set containing 28% of noisy patterns uniformly added to all classes. (b) CS^{4}VM. (c) SVM. (d) *k*-NN. (e) ML.

4.5.2 Results with mislabeled training patterns concentrated on a specific class

In the second set of experiments, several samples of the class “forest” were added to the class “fields” to reach 10%, 16%, 22%, 28% of mislabeled patterns with respect to the total number of training samples. Also in this case the presence of errors that systematically affected one class severely impacted the performance of the supervised classification algorithms. When a low percentage (10%) of noisy patterns was added to the original training set, all the considered classifiers decreased their kappa coefficient of accuracy by more than 12% (see Table 4.9 and Fig. 4.14). In contrast to the first set of experiments, also the SVM algorithm suffered the presence of this type of noisy training set, reducing its accuracy by 18.4% (while the *k*-NN decreased its accuracy by 20.2% and the ML by 22.5%). The semisupervised approach based on the PS^{3}VM was not able to improve the accuracies of the standard SVM. The CS^{4}VM could partially recover the accuracy of standard SVM by increasing the kappa accuracy by 7.4%, thus limiting the effect of mislabeled patterns. When the amount of noisy patterns further increased, PS^{3}VM, SVM, ML and *k*-NN classifiers did not further decrease significantly their kappa accuracies.

Table 4.9 - Kappa coefficient of accuracy on the test set using training sets with different percentages of mislabeled patterns added to a specific class (Landsat data set).

% of mislabeled patterns	**Kappa Accuracy**				
	CS^{4}VM	**PS^{3}VM**	**SVM**	***k*-NN**	**ML**
0	0.927	0.915	0.919	0.882	0.923
10	0.809	0.738	0.735	0.680	0.699
16	0.712	0.706	0.695	0.652	0.678
22	0.691	0.664	0.661	0.632	0.671
28	0.658	0.651	0.648	0.632	0.666

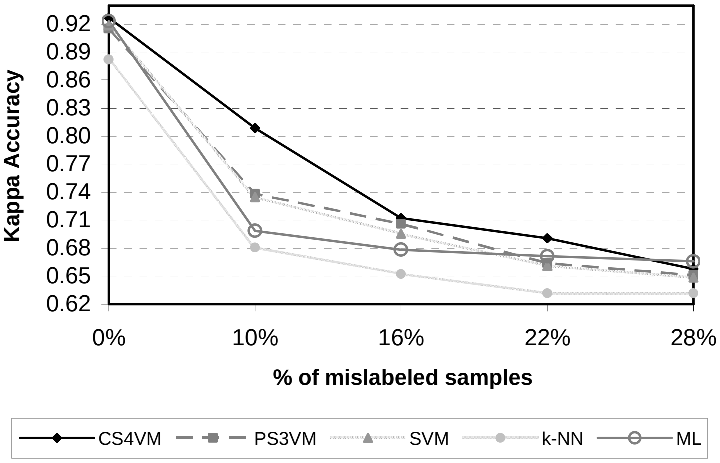


Fig. 4.14 - Behavior of the kappa coefficient of accuracy on test set versus the percentage of mislabeled training patterns concentrated on a specific class (Landsat data set).

This behavior is confirmed from the average kappa coefficient of accuracy of the binary classifiers versus the percentage of mislabeled training patters (see Fig. 4.15). In this case we do not report the results of the binary classifiers exhibiting the lowest accuracy because the complexity of the problem resulted in unreliable kappa values on this class (even if also in this case the CS^{4}VM outperformed the SVM).

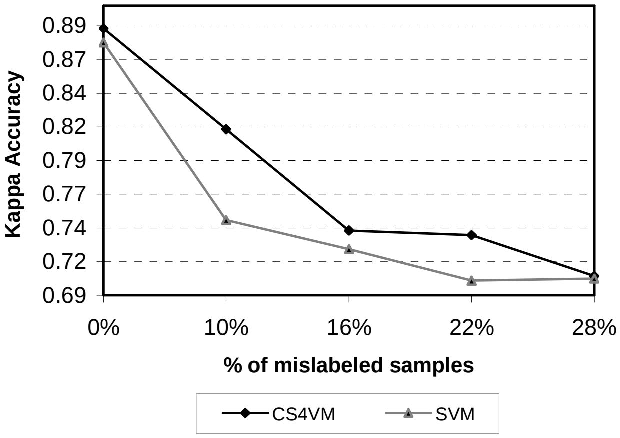


Fig. 4.15 - Behavior of the average kappa coefficient of accuracy (computed on all the binary CS^{4}VMs and SVMs included in the multiclass architecture) versus the percentage of mislabeled training patterns concentrated on a specific class (Landsat data set).

4.6 Discussion and conclusion

In this chapter we have proposed a novel classification technique based on SVM that exploits the contextual information in order to render the learning of the classifier more robust to possible mislabeled patterns present in the training set. Moreover, we have analyzed the effects of mislabeled training samples on the classification accuracy of supervised algorithms, comparing the results obtained by the proposed CS^{4}VM with those yielded by a progressive semisupervised SVM (PS^{3}VM), a standard supervised SVM, a Gaussian ML, and a *k*-NN. This analysis was carried out varying both the percentage of mislabeled patterns and their distribution on the information classes. The experimental results obtained on two different data sets (a VHR image acquired by the Ikonos satellite and a medium resolution image acquired by the Landsat 5 satellite) confirm that the proposed CS^{4}VM approach exhibits augmented robustness to noisy training sets with respect to all the other classifiers. In greater detail, the proposed CS^{4}VM method always increased the average kappa coefficient of accuracy of the binary classifiers included in the OAA multiclass architecture with respect to the standard SVM classifier. Moreover, in many cases the CS^{4}VM sharply increased the accuracy on the information class that was most affected by the mislabeled patterns introduced in the training set.

By analyzing the effects of the distribution of mislabeled patterns on the classes, it is possible to conclude that errors concentrated on a class (or on a subset of classes) are much more critical than errors uniformly distributed on all classes. In greater detail, when noisy patterns were added uniformly to all classes, we observed that the proposed CS^{4}VM resulted in higher and more stable accuracies than all the other classifiers. The supervised SVM and the PS^{3}VM exhibited relatively high accuracies when a moderate amount of noisy patterns was included in the training set, but they slowly decreased their accuracy when the percentage of mislabeled samples increased. On the contrary, both the ML and the *k*-NN classifiers are very sensitive even to the presence of a small amount of noisy patterns, and sharply decreased their accuracies by increasing the number of mislabeled samples. Nevertheless, the *k*-NN classifier resulted significantly more accurate than the ML classifier when mislabeled patterns equally affected the considered information classes. When noisy patterns were concentrated on a specific class of the training set, the accuracies of all the considered classifiers sharply decreased by increasing the amount of mislabeled training samples. Moreover, in this case, the proposed CS^{4}VM exhibited, in general, the highest and more stable accuracies. Nonetheless, when the number of mislabeled patterns increased over a given threshold, the classification problem became very critical and also the proposed technique significantly reduced its effectiveness. The standard SVM classifier still maintained higher accuracies than the ML and the *k*-NN techniques. The PS^{3}VM slightly increased the accuracies of the standard SVM. Unlike the previous case, the *k*-NN algorithm resulted in lower accuracies than the ML method. This is mainly due to the fact that mislabeled patterns concentrated on a single class (or on few classes) alter the prior probabilities, thus affecting more the *k*-NN classifier (which implicitly considers the prior probabilities in the decision rule) than the ML technique (which does not consider the prior probabilities of classes).

The proposed CS^{4}VM introduces some additional free parameters with respect to the standard supervised SVM, which should be tuned in the model-selection phase. The analysis on the effects of the values of these parameters on the classification results (carried out in the different simulations described in this chapter) pointed out that the empirical selection of $\mathrm{K} = \kappa_1 / \kappa_2 = 2$ (which is reasonable considering the physical meaning of this ratio) resulted in good accuracies on both data sets. This choice allows one to reduce the model-selection phase to tune the value of the ratio C/κ_1 in addition to the standard SVM parameters. Nonetheless, when possible, the inclusion of the choice of the κ_1 / κ_2 value in the model selection would optimize the results achievable with the proposed approach. The optimal value for the ratio C/κ_1 depends on the considered data set and the type of mislabeling errors, but in general we observed that higher weights for the context patterns (lower values for the ratio C/κ_1) can result in better classification accuracies when the percentage of mislabeled training patterns increases. This confirms the importance of the context term to increase the classification accuracy in presence of noisy training sets.

It is worth noting that the considered PS^{3}VM classifier slightly improved the accuracy with respect to the standard SVM by exploiting the information of unlabeled samples, but it could not gain in accuracy when the amount of mislabeled patterns increased. Indeed, the PS^{3}VM is not developed to take into account the possible presence of mislabeled training patterns, which affect the first iteration of the learning phase propagating the errors to the semilabeled samples in the next iterations of the algorithm. On the contrary, the proposed CS^{4}VM is especially developed to cope with “non fully reliable” training sets by exploiting the information of pixels in the neighborhood of the training points according to a specific weighting mechanism that penalizes less reliable training patterns. In addition, the proposed CS^{4}VM approach is computationally less demanding than the PS^{3}VM as it requires only two steps (this choice is done for limiting the computational complexity and is supported from empirical experiments that confirmed that increasing the number of iterations does not significantly change the classification results). On the contrary, the PS^{3}VM may require a large number of iterations before convergence.

The computational cost of the learning phase of the proposed CS^{4}VM method is slightly higher than that required from the standard supervised SVM. This depends on both the second step of the learning algorithm (which involves an increased number of samples, as semilabeled context patterns are considered in the process) and the setting of the additional parameters in the model-selection phase. In our experiments on the Ikonos data set, carried out on a PC mounting an Intel Pentium D processor at 3.4 GHz and a 2-Gb DDR2 RAM, the training phase of a supervised SVM took in average about 20 seconds, while the one of the proposed CS^{4}VM required about 3 minutes. It is important to point out that the additional cost of the proposed method concerns only the learning phase, whereas the computational time in the classification phase remains unchanged.

As a final remark, it is worth stressing that proposed analysis points out the dramatic effects involved on the classification accuracy from a relatively small percentages of mislabeled training samples concentrated on a class (or on a subset of classes). This should be understood in order to define adequate strategies in the design of training data for avoiding this kind of errors.

4.7 References

[1] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, 2nd ed., New York: Wiley, 2001.

[2] J. A. Richards, X. Jia, Remote Sensing Digital Image Analysis, 4th ed., Berlin, Germany: Springer-Verlag, 2006.

[3] M. Chi and L. Bruzzone, “An ensemble-driven k-NN approach to ill posed classification problems,” *Pattern Recognition Letters*, vol. 27, no. 4, pp. 301–307, Mar. 2006.

[4] V. N. Vapnik, The Nature of Statistical Learning Theory, 2nd ed., New York: Springer, 2001.

[5] C. Burges, “A tutorial on support vector machines for pattern recognition,” Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 121–167, 1998.

[6] N. Cristianini and J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods, Cambridge, U.K.: University press, 1995.

[7] B. Schölkopf and A. Smola, Learning With Kernels, Cambridge, MA: MIT Press, online: http://www.learning-with-kernels.org/, 2002.

[8] F. Melgani, L. Bruzzone, “Classification of Hyperspectral Remote Sensing Images With Support Vector Machines,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778- 1790, Aug. 2004.

[9] G. F. Hughes, “On the mean accuracy of statistical pattern recognition,” *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 55–63, Jan. 1968.

[10] F. Bovolo, L. Bruzzone, “A Context-Sensitive Technique Based on Support Vector Machines for Image Classification,” *IEEE Pattern Recognition and Machine Intelligence Conference (PReMI 2005)*, Lecture Notes in Computer Science, vol. 3776, Kolkata-India, Dec. 2005.

[11] A. A. Farag, R. M. Mohamed, A. El-Baz, “A unified framework for MAP estimation in remote sensing image segmentation,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 7, pp. 1617-1634, July 2005.

[12] F. Melgani and S. Serpico, “A Markov Random Field Approach to Spatio-Temporal Contextual Image Classification”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 11, pp. 2478–2487, Nov. 2003.

[13] G. Moser, S. Serpico, F. Causa, “MRF model parameter estimation for contextual supervised classification of remote-sensing images”, in *Proc. IEEE International Geoscience and Remote Sensing Symposium*, (IGARSS '05), pp. 308-311, July 2005.

[14] P. Gamba, F. Dell’Acqua, G. Lisini, and G. Trinni, “Improved VHR Urban Area Mapping Exploiting Object Boundaries”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 8, pp. 2676-2682, Aug. 2007.

[15] M. Berthod, Z. Kato, S. Yu, and J. Zerubia, “Bayesian Image Classification Using Markov Random Fields”, Image and Vision Computing, vol. 14, pp. 285-295, 1996.

[16] R. Nishii, “A Markov Random Field-Based Approach to Decision-Level Fusion for Remote Sensing Image Classification”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 10, pp. 2316-2319, Oct. 2003.

[17] L. Bruzzone, M. Chi and M. Marconcini, Semisupervised support vector machines for classification of hyperspectral remote sensing images, chapter 11, Hyperspectral data explotation, Chein-I Chang, Wiley, USA, pp.275-311, 2007.

[18] L. Bruzzone, M. Chi, M. Marconcini, “A Novel Transductive SVM for Semisupervised Classification of Remote-Sensing Images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 11, Part 2, pp. 3363-3373, Nov. 2006.

[19] M. M. Dundar and D. A. Landgrebe, “A cost-effective semisupervised classifier approach with kernels,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 1, pp. 264–270, Jan. 2004.

[20] K. P. Bennett and A. Demiriz, “Semi-supervised support vector machines,” *Advances in Neural Information Processing Systems*, vol. 10. Cambridge, MIT Press, pp. 368–374, 1998.

[21] C. Hsu and C. Lin, “A comparison of methods for multi-class support vector machines,” *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[22] R. G. Congalton, K. Green, Assessing the Accuracy of Remotely Sensed Data, Boca Raton, U.S.A: Lewis Publishers, 1999.

[23] J. Platt, Fast training of support vector machines using sequential minimal optimization, chapter 12, Advances in Kernel Methods: Support Vector Learning, B. Schölkopf, C. Burges, and A. Smola, Cambridge, MA: MIT Press, pp. 185-208, 1998.

[24] B. Aiazzi, S. Baronti, M. Selva, L. Alparone, “Enhanced Gram-Schmidt Spectral Sharpening Based on Multivariate Regression of MS and Pan Data,” *in Proc. IEEE International Geoscience and Remote Sensing Symposium*, (IGARSS '06), pp. 3806-3809, 2006.

Chapter 5

5. Batch Mode Active Learning Methods for the Interactive Classification of Remote Sensing Images

This chapter investigates different batch mode active learning techniques for the classification of remote sensing (RS) images with support vector machines (SVMs). This is done by generalizing to multiclass problems techniques defined for binary classifiers. The investigated techniques exploit different query functions, which are based on the evaluation of two criteria: uncertainty and diversity. The uncertainty criterion is associated to the confidence of the supervised algorithm in correctly classifying the considered sample, while the diversity criterion aims at selecting a set of unlabeled samples that are as more diverse (distant one another) as possible, thus reducing the redundancy among the selected samples. The combination of the two criteria results in the selection of the potentially most informative set of samples at each iteration of the active learning process. Moreover, we propose a novel query function that is based on a kernel clustering technique for assessing the diversity of samples and a new strategy for selecting the most informative representative sample from each cluster. The investigated and proposed techniques are theoretically and experimentally compared with state-of-the-art methods adopted for RS applications. This is accomplished by considering VHR multispectral and hyperspectral images. By this comparison we observed that the proposed method resulted in better accuracy with respect to other investigated and state-of-the art methods on both the considered data sets. Furthermore, we derived some guidelines on the design of active learning systems for the classification of different types of RS images.

5.1 Introduction

Land cover classification from RS images is generally performed by using supervised classification techniques, which require the availability of labeled samples for training the supervised

This chapter was submitted to the *IEEE Transactions on Geoscience and Remote Sensing*. Title: "Batch Mode Active Learning Methods for the Interactive Classification of Remote Sensing Images". Authors: B. Demir, C. Persello, and L. Bruzzone.

algorithm. As we observed in the previous chapter, the amount and the quality of the available training samples are crucial for obtaining accurate classification maps. However, the collection of labeled samples is time consuming and costly, and the available training samples are often not enough for an adequate learning of the classifier. A possible approach to address this problem is to exploit unlabeled samples in the learning of the classification algorithm according to semisupervised or transductive classification procedure. The semisupervised approach has been widely investigated in the recent years in the RS community [3]-[5]. A different approach to both enrich the information given as input to the supervised classifier and improve the statistic of the classes is to iteratively expand the original training set according to a process that requires an interaction between the user and the automatic recognition system. This approach is known in the machine learning community as active learning (AL) and, although marginally considered in the RS community, can result very useful for different applications. The AL process is conducted according to an iterative process. At each iteration, the most informative unlabeled samples are chosen for a manual labeling and the supervised algorithm is retrained with the additional labeled samples. In this way, the unnecessary and redundant labeling of non informative samples is avoided, greatly reducing the labeling cost and time. Moreover, AL allows one to reduce the computational complexity of the training phase. In this chapter we focus our attention on AL methods.

In RS classification problems, the collection of labeled samples for the initial training set and the labeling of queried samples can be derived according to: 1) in situ ground surveys (which are associate to high cost and require time), or 2) image photointerpretation (which is cheap and fast). The choice of the labeling strategy depends on the considered problem and image. For example, we can reasonably suppose that for the classification of very high resolution (VHR) images, the labeling of samples can be easily carried out by photointerpretation. Indeed, the metric or sub-metric resolution of VHR images allows a human expert to identify and label the objects on the ground and the different land-cover types on the basis of the inspection of real or false color compositions. On the contrary, when medium (or low) resolution multispectral images and hyperspectral data are considered, ground surveys are usually required. Medium and low resolution images do not usually allow one to recognize the objects on the ground, and the land-cover classes of the pixels (which may be associated to different materials) cannot usually be recognized with high reliability by a human expert. Hyperspectral data, thanks to a dense sampling of the spectral signature, allows one characterizing several different land-cover classes (e.g., associated to different arboreal species) that cannot be recognized by a visual analysis of different false color compositions. Thus, depending on both the type of classification problem and the considered type of data, the cost and time associated to the labeling process significantly changes. These different scenarios require the definition of different AL schemes: we expect that in cases where photointerpretation is possible, several iterations of the labeling step may be carried out; whereas in cases where ground truth surveys are necessary, only few iterations (e.g., two or three) of the AL process are possible.

Most of the previous studies in AL have focused on selecting the single most informative sample at each iteration, by assessing its uncertainty [6]-[12]. This can be inefficient, since the classifier has to be retrained for each new labeled sample. Moreover, this approach is not appropriate for RS image classification tasks for the abovementioned reasons (both in the case of photointerpretation and ground surveys for sample labeling). Thus, in this chapter we focus on batch mode active learning, where a batch of $h > 1$ unlabeled samples is queried at each iteration. The problem with such an approach is that by selecting the samples of the batch on the basis of the uncertainty only, some of the selected samples could be similar to each other, and thus do not provide additional information for the model updating with respect to other samples in the batch. The key issue of batch mode AL is to select sets of samples with little redundancy, so that they can provide the highest possible information to the classifier. Thus, the query function adopted for selecting the batch of the most informative samples should take into account two main criteria: 1) uncertainty, and 2) diversity of samples [13]-[15]. The uncertainty criterion is associated to the confidence of the supervised algorithm in correctly classifying the considered sample, while the diversity criterion aims at selecting a set of unlabeled samples that are as more diverse (distant one another) as possible, thus reducing the redundancy among the selected samples. The combination of the two criteria results in the selection of the potentially most informative set of samples at each iteration of the AL process.

The aim of this chapter is to investigate different AL techniques proposed in the machine learning literature and to properly generalize them to the classification of RS images with multiclass problem addressed by support vector machines (SVMs). The investigated techniques use different query functions with different strategies to assess the uncertainty and diversity criteria in the multiclass case. Moreover, we propose a novel query function that is based on a kernel clustering technique for assessing the diversity of samples and a new strategy for selecting the most informative representative sample from each cluster. The investigated and proposed techniques are theoretically and experimentally compared among them and with other AL algorithms proposed in the RS literature in the classification of VHR images and hyperspectral data. On the basis of this comparison some guidelines are derived on the use of AL techniques for the classification of different types of RS images.

The rest the chapter is organized as follows. Section 5.2 reviews the background on AL methods and their application to RS problems. Section 5.3 presents the investigated batch mode AL techniques and the proposed generalization to multiclass problems. Section 5.4 presents the proposed novel query function based on kernel clustering and an original selection of cluster most informative samples. Section 5.5 presents the description of the two considered VHR and hyperspectral data sets and the design of experiments. Section 5.6 illustrates the results obtained by the extensive experimental analysis carried out on the considered data sets. Finally, Section 5.7 draws the conclusion of this chapter.

5.2 Background on active learning

5.2.1 Active learning process

A general active learner can be modeled as a quintuple (*G*, *Q*, *S*, *T*, *U*) [6]. G is a supervised classifier, which is trained on the labeled training set *T*. *Q* is a query function used to select the most informative unlabeled samples from a pool *U* of unlabeled samples. *S* is a supervisor who can assign the true class label to any unlabeled sample of *U*. The AL process is an iterative process, where the supervisor *S* interacts with the system by iteratively labeling the most informative samples selected by the query function *Q* at each iteration. At the initial stage, an initial training set *T* of few labeled samples is required for the first training of the classifier *G*. After initialization, the query function *Q* is used to select a set of samples *X* from the pool *U* and the supervisor

S assigns them the true class label. Then, these new labeled samples are included into *T* and the classifier *G* is retrained using the updated training set. The closed loop of querying and retraining continues for some predefined iterations or until a stop criterion is satisfied. Algorithm 1 gives a description of a general AL process.

Algorithm 1: Active learning procedure

```
1. Train the classifier $G$ with the initial training set $T$
2. Classify the unlabeled samples of the pool $U$
Repeat
3. Query a set of samples (with query function $Q$) from the pool $U$
4. A label is assigned to the queried samples by the supervisor $S$
5. Add the new labeled samples to the training set $T$
6. Retrain the classifier
Until a stopping criteria is satisfied.
```

The query function *Q* is of fundamental importance in AL techniques, which often differ only in their query functions. Several methods have been proposed so far in the machine learning literature. A probabilistic approach to AL is presented in [7], which is based on the estimation of the posterior probability density function of the classes both for obtaining the classification rule and to estimate the uncertainty of unlabeled samples. In the two-class case, the query of the most uncertain samples is obtained by choosing the samples closest to 0.5 (half of them below and half above this probability value). The query function proposed in [16] is designed to minimize future errors, i.e., the method selects the unlabeled pattern that, once labeled and added to the training data, is expected to result in the lowest error on test samples. This approach is applied to two regression models (i.e., weighted regression and mixture of Gaussians) where an optimal solution for minimizing future error rates can be obtained in closed form. Unfortunately, this solution is intractable to calculate the expected error rate for most classifiers without specific statistical models. A statistical learning approach is also used in [17] for regression problems with multilayer perceptron. In [18], a method is proposed that selects the next example according to an optimal criterion (which minimizes the expected error rate on future test samples), but solves the problem by using a sampling estimation. Two methods for estimating future error rate are presented. In the first method, the future error rate is estimated by log-loss using the entropy of the posterior class distribution on the set of unlabeled samples. In the second method, a 0-1 loss function using the posterior probability of the most probable class for a set of unlabeled samples is used.

Another popular paradigm is given by committee-based active learners. The “query by committee” approach [19]-[21] is a general AL algorithm that has theoretical guarantees on the reduction in prediction error with the number of queries. A committee of classifiers using different hypothesis about parameters is trained to label a set of unknown examples. The algorithm selects the samples where the disagreement between the classifiers is maximal. In [22], two query methods are proposed that combine the idea of query by committee and that of boosting and bagging.

An interesting category of AL approaches, which have gained significant success in numerous real-world learning tasks, is based on the use of support vector machines (SVMs) [8]-[14]. The SVM classifier [4]-[8] is particularly suited to AL due to its intrinsic high generalization capabilities and because its classification rule can be characterized by a small set of support vectors that can be easily updated over successive learning iterations [12]. One of the most popular (and effective) query heuristic for active SVM learning is to select the data point closes to the current separating hyperplane, which is also referred to as margin sampling (MS). This method results in the selection of the unlabeled sample with the lowest confidence, i.e., the maximal uncertainty on the true information class. The query strategy proposed in [10] is based on the splitting of the version space [10],[13]: the point which split the current version space into two halves having equal volumes are selected at each step, as they are likely to be the actual support vectors. Three heuristics for approximating the above criterion are described, the simplest among them selects the point closes to the hyperplane as in [8]. In [6], an approach is proposed that estimates the uncertainty level of each sample according to the output score of a classifier and selects only those samples whose outputs are within the uncertainty range. In [11], the authors present possible generalizations of the active SVM approach to multiclass problems.

It is important to observe that the abovementioned methods consider only the uncertainty of samples, which is an optimal criterion only for the selection of one sample at each iteration. Selecting a batch of $h > 1$ samples exclusively on the basis of the uncertainty (e.g., the distance to the classification hyperplane) may result in the selection of similar (redundant) samples that do not provide additional information. However, in many problems it is necessary to speed up the learning process by selecting batches of more than one sample at each iteration. In order to address this shortcoming, in [13] an approach is presented especially designed to construct batches of samples by incorporating a diversity measure that considers the angles between the induced classification hyperplanes (more details on this approach are given in the next section). Another approach to consider the diversity in the query function is the use of clustering [14]-[15]. In [14], an AL heuristic is presented, which explores the clustering structure of samples and identifies uncertain samples avoiding redundancy (details of this approach are given in the next section). In [25]-[26], the authors present a framework for batch mode AL that applies the Fisher information matrix to select a number of informative examples simultaneously.

Nevertheless, most of the abovementioned approaches are designed for binary classification and thus are not suitable for most of the RS classification problems. In this chapter, we focus on multiclass SVM-based AL approaches that can select a batch of samples at each iteration for the classification of RS images. The next subsection provides a discussion and a review on the use of AL for the classification of RS images.

5.2.2 Active learning for the classification of RS data

Active learning has been applied mainly to text categorization and image retrieval problems. However, the AL approach can be adopted for the interactive classification of RS images by taking into account the peculiarities of this domain. In RS problems, the supervisor S is a human expert that can derive the land-cover type of the area on the ground associated to the selected patterns according to the two possible strategies identified in the introduction, i.e., photointerpretation and ground survey. These strategies are associated with significantly different costs. It is important to note that the use of photointerpretation or of ground surveys (and thus the cost) depends on the considered classification problem, i.e., the type of the considered RS image, and the set of land-cover classes. Moreover, the cost of ground surveys also depends on the considered geographical area. In [27], the AL problem is formulated considering a spatially dependent label acquisition costs. In the present work we consider that the labeling cost mainly depends on the type of the RS data, which affects the aforementioned labeling strategy. For example, in case of VHR images, often the labeling of samples can be carried out by photointerpretation, while in the case of medium/low resolution multispectral images and hyperspectral data, ground surveys are necessary. No particular restrictions are usually considered for the definition of the initial training set *T*, since we expect that the AL process can be started up with few samples for each class without affecting the convergence capability (the initial samples can affect the number of iterations necessary for obtaining convergence). The pool of unlabeled samples *U* can be associated to the whole considered image or to a portion of it (for reducing the computational time associated to the query function and/or for considering only the areas of the scene accessible for labeling). An important issue is related to the capability of the query function to select batches of $h > 1$ samples, which results to be of fundamental importance for the adoption of AL in real-world RS problems. It is worth to stress here the importance of the choice of the *h* value in the design of the AL classification system, as it affects the number of iterations and thus both the performance and the cost of the classification system. In general, we expect that for the classification of VHR images (where photointerpretation is possible), several iterations of the labeling step may be carried out and small values for *h* can be adopted; whereas in cases where ground truth surveys are necessary, only few iterations (e.g., two or three) of the AL process are possible and large *h* values are necessary.

In the RS domain, AL was applied to the detection of subsurface targets, such as landmines and unexploded ordnance in [29]-[30]. Some preliminary works about the use of AL for RS classification problems can be found in [12], [31]-[32]. The technique proposed in [12] is based on MS and selects the most uncertain sample for each binary SVM in a OAA multiclass architecture (i.e., querying $h = L$ samples, where *L* is the number of classes). In [31], two batch mode AL techniques for multiclass RS classification problems are proposed. The first technique is MS by closest support vector (MS-cSV), which considers the smallest distance of the unlabeled samples to the *L* hyperplanes (associated to the *L* binary SVMs in a OAA multiclass architecture) as the uncertainty value. At each iteration, the most uncertain unlabeled samples, which do not share the closest SV, are added to the training set. The second technique, called entropy query-by bagging (EQB), is based on the selection of unlabeled samples according to the maximum disagreement between a committee of classifiers. The committee is obtained by bagging: first different training sets (associated to different EQB predictors) are drawn with replacement from the original training data. Then, each training set is used to train the OAA SVM architecture to predict the different labels for each unlabeled sample. Finally, the entropy of the distribution of the different labels associated to each sample is calculated to evaluate the disagreement among the classifiers on the unlabeled samples. The samples with maximum entropy (i.e., those with maximum disagreement among the classifiers) are added to the current training set. In [32], an AL technique is presented, which selects the unlabeled sample that maximizes the information gain between the a posteriori probability distribution estimated from the current training set and the training set obtained by including that sample into it. The information gain is measured by the Kullback–Leibler (KL) divergence. This KL-Maximization (KL-Max) technique can be implemented with any classifier that can estimate the posterior class probabilities. However this technique can be used to select only one sample at each iteration.

5.3 Investigated query functions

In this section we investigate different query functions Q based on SVM for multiclass RS classification problems. The investigated techniques are based on standard methods; however, some of them are presented here with modifications with respect to the original version to overcome shortcomings that would affect their applicability to real RS problems. In particular, the presented techniques are adapted to classification problems characterized by a number of classes $L > 2$ (multiclass problems) and to the inclusion of a batch of $h > 1$ samples at each iteration in the training set (for taking into account RS constraints and limiting the AL process to few iterations according to the analysis presented in the previous sections). The investigated query functions are based on the evaluation of the uncertainty and diversity criteria applied in two consecutive steps. The $m > h$ most uncertain samples are selected in the uncertainty step and the most diverse h ($h > 1$) samples among these m uncertain samples are chosen in the diversity step. The ratio m / h provides an indication on the tradeoff between uncertainty and diversity. In this section we present different possible implementations for both steps, focusing on the OAA multiclass architecture.

5.3.1 Techniques for implementing the uncertainty criterion with multiclass SVMs

The uncertainty criterion aims at selecting the samples that have maximum uncertainty among all samples in the unlabeled sample pool *U*. Since the most uncertain samples have the lowest probability to be correctly classified, they are the most useful to be included in the training set. In this chapter, we investigate two possible techniques in the framework of multiclass SVM: a) binary-level uncertainty (which evaluates uncertainty at the level of binary SVM classifiers), and b) multiclass-level uncertainty (which analysis uncertainty within the considered OAA architecture).

Binary-level uncertainty (BLU)

The binary-level uncertainty (BLU) technique separately selects a batch of the most uncertain unlabeled samples from each binary SVM on the basis of the MS query function. In the technique adopted in [12], only the unlabeled sample closest to the hyperplane of each binary SVM was added to the training set at each iteration (i.e., $h = L$). On the contrary, in the investigated BLU technique, at each iteration the most uncertain q ($q > 1$) samples are selected from each binary SVM (instead of a single sample). In greater detail, *L* binary SVMs are initially trained with the current training set and the functional distance $f_i(\mathbf{x})$, $i = 1,...,L$ of each unlabeled sample $\mathbf{x} \in U$ to the hyperplane is obtained. Then, the set of q samples $\left\{\mathbf{x}_{1,i}^{BLU}, \mathbf{x}_{2,i}^{BLU}, ..., \mathbf{x}_{q,i}^{BLU}\right\}$, $i = 1, 2, ..., L$ closest to margin of the corresponding hyperplane are selected for each binary SVM. Totally $\rho = qL$ samples are taken. Note that $\mathbf{x}_{j,i}^{BLU}$, $j = 1, 2, ..., q$, represents the selected j-th sample from the *i*-th SVM. Since some unlabeled samples can be selected by more than one binary SVM, the redundant samples are removed. Thus, the total number *m* of selected samples can actually be smaller than ρ (i.e., $m \le \rho$). The set of m most uncertain samples $\{\mathbf{x}_1^{BLU}, \mathbf{x}_2^{BLU}, ..., \mathbf{x}_m^{BLU}\}$ is forwarded to the diversity step. Fig. 5.1 shows the architecture of the investigated BLU technique.

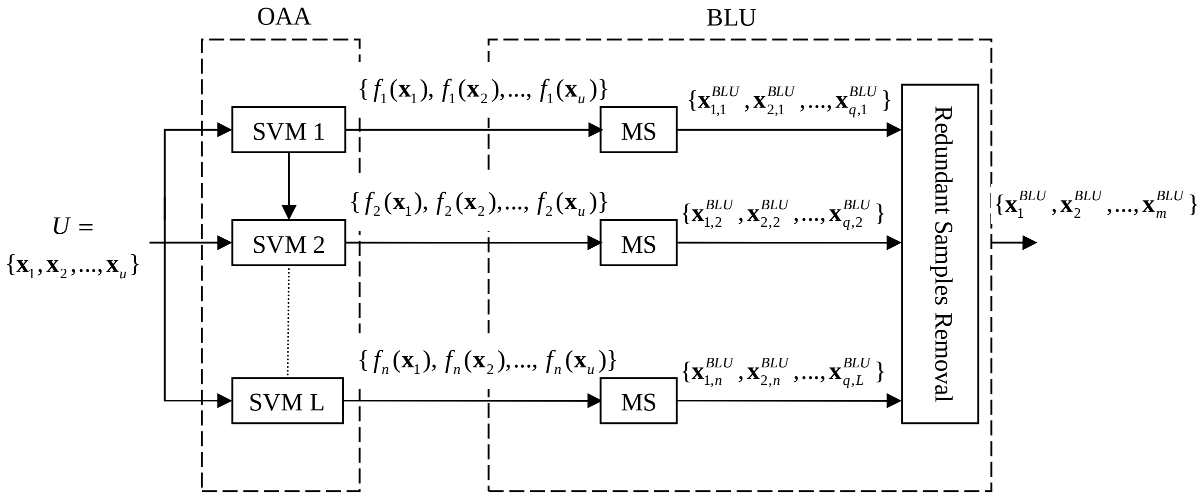

OAA
BLU
$U = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_u\}$
SVM 1
SVM 2
SVM L
$\{f_1(\mathbf{x}_1), f_1(\mathbf{x}_2), ..., f_1(\mathbf{x}_u)\}$
$\{f_2(\mathbf{x}_1), f_2(\mathbf{x}_2), ..., f_2(\mathbf{x}_u)\}$
$\{f_n(\mathbf{x}_1), f_n(\mathbf{x}_2), ..., f_n(\mathbf{x}_u)\}$
MS
MS
MS
$\{\mathbf{x}_{1,1}^{BLU}, \mathbf{x}_{2,1}^{BLU}, ..., \mathbf{x}_{q,1}^{BLU}\}$
$\{\mathbf{x}_{1,2}^{BLU}, \mathbf{x}_{2,2}^{BLU}, ..., \mathbf{x}_{q,2}^{BLU}\}$
$\{\mathbf{x}_{1,n}^{BLU}, \mathbf{x}_{2,n}^{BLU}, ..., \mathbf{x}_{q,L}^{BLU}\}$
Redundant Samples Removal
$\{\mathbf{x}_1^{BLU}, \mathbf{x}_2^{BLU}, ..., \mathbf{x}_m^{BLU}\}$

Fig. 5.1 - Multiclass architecture adopted for the BLU technique

Multiclass-level uncertainty (MCLU)

The adopted multiclass-level uncertainty (MCLU) technique selects the most uncertain samples according to a confidence value $c(\mathbf{x})$, $\mathbf{x} \in U$, which is defined on the basis of their functional distance $f_i(\mathbf{x})$, $i = 1,...,L$ to the L decision boundaries of the binary SVM classifiers included in the OAA architecture [31], [33]. In this technique, the distance of each sample $\mathbf{x} \in U$ to each hyperplane is calculated and a set of L distance values $\{f_1(\mathbf{x}), f_2(\mathbf{x}),... f_n(\mathbf{x})\}$ is obtained. Then, the confidence value $c(\mathbf{x})$ can be calculated using different strategies. Here, we consider two strategies: 1) the minimum distance function $c_{\min}(\mathbf{x})$ strategy, which is obtained by taking the smallest distance to the hyperplanes (as absolute value), i.e., [31]

$$c_{\min}(\mathbf{x}) = \min_{i=1,2,...,n} \{abs[f_i(\mathbf{x})]\} \tag{5.1}$$

and 2) the difference $c_{diff}(\mathbf{x})$ strategy, which considers the difference between the first largest and the second largest distance values to the hyperplanes (note that, for the ***i***-th binary SVM in the OAA architecture, $f_i(\mathbf{x}) \geq 0$ if x belongs to ***i***-th class and $f_i(\mathbf{x}) < 0$ if x belongs to the rest), i.e, [33]

$$\begin{aligned} r_{1\max} &= \arg\max_{i=1,2,...,n} \{f_i(\mathbf{x})\} \\ r_{2\max} &= \arg\max_{j=1,2,...,n,\ j\neq r_{1\max}} \{f_j(\mathbf{x})\} \\ c_{diff}(\mathbf{x}) &= f_{r_{1\max}}(\mathbf{x}) - f_{r_{2\max}}(\mathbf{x}) \end{aligned} \tag{5.2}$$

The $c_{\min}(\mathbf{x})$ function models a simple strategy that computes the confidence of a sample **x** taking into account the minimum distance to the hyperplanes evaluated on the basis of the most uncertain binary SVM classifier. Differently, the $c_{diff}(\mathbf{x})$ strategy assesses the uncertainty between the two most likely classes. If this value is high, the sample **x** is assigned to $r_{1\max}$ with high confidence. On the contrary, if $c_{diff}(\mathbf{x})$ is small, the decision for $r_{1\max}$ is not reliable and there is a possible conflict with the class $r_{2\max}$ (i.e., the sample **x** is very close to the boundary between class $r_{1\max}$ and $r_{2\max}$). Thus, this sample is considered uncertain and is selected by the query function for better modeling the decision function in the corresponding position of the feature space. After that the $c(\mathbf{x})$ value of each $\mathbf{x} \in U$ is obtained based on one of the two above-mentioned strategies, the m samples $\mathbf{x}_1^{MCLU}, \mathbf{x}_2^{MCLU}, ..., \mathbf{x}_m^{MCLU}$ with lower $c(\mathbf{x})$ are selected to be forwarded to the diversity step. Note that $\mathbf{x}_j^{MCLU}$ denotes the selected j-th most uncertain sample based on the MCLU strategy. Fig. 5.2 shows the architecture of the investigated MCLU technique.

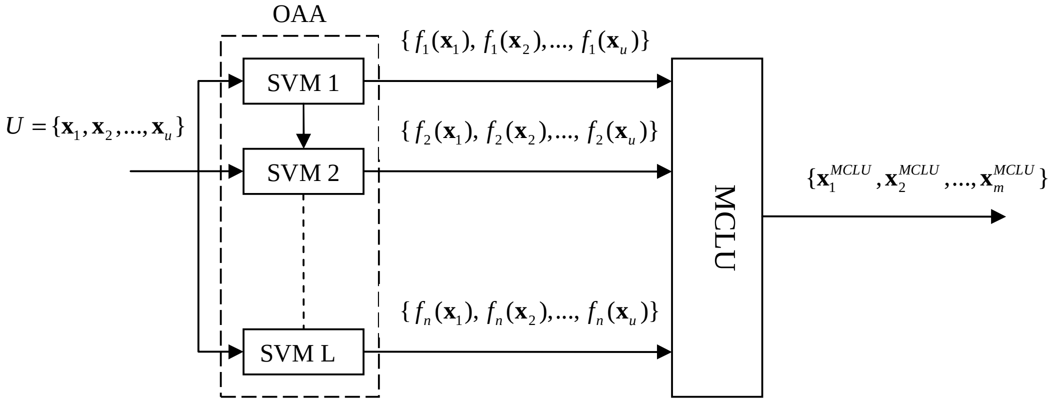

OAA
$U = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_u\}$
SVM 1
SVM 2
SVM L
$\{f_1(\mathbf{x}_1), f_1(\mathbf{x}_2), ..., f_1(\mathbf{x}_u)\}$
$\{f_2(\mathbf{x}_1), f_2(\mathbf{x}_2), ..., f_2(\mathbf{x}_u)\}$
$\{f_n(\mathbf{x}_1), f_n(\mathbf{x}_2), ..., f_n(\mathbf{x}_u)\}$
MCLU
$\{\mathbf{x}_1^{MCLU}, \mathbf{x}_2^{MCLU}, ..., \mathbf{x}_m^{MCLU}\}$

Fig. 5.2 - Architecture adopted for the MCLU technique.

5.3.2 Techniques for implementing the diversity criterion

The main idea of using diversity in AL is to select a batch of samples ($h > 1$) which have low confidence values (i.e., the most uncertain ones), and at the same time are diverse from each other. In this chapter, we consider two diversity methods: 1) the angle based diversity (ABD); and 2) the clustering based diversity (CBD). Before considering the multiclass formulation, in the following we recall their definitions for two-class problems.

Angle based diversity (ABD)

A possible way for measuring the diversity of uncertain samples is to consider the cosine angle distance. It is a similarity measure between two samples defined in the kernel space by [13]

$$\begin{aligned} \left|\cos\left(\angle(\mathbf{x}_i, \mathbf{x}_j)\right)\right| &= \frac{\left|\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)\right|}{\left\|\phi(\mathbf{x}_i)\right\| \left\|\phi(\mathbf{x}_j)\right\|} = \frac{K(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{K(\mathbf{x}_i, \mathbf{x}_i) K(\mathbf{x}_j, \mathbf{x}_j)}} \\ \angle(\mathbf{x}_i, \mathbf{x}_j) &= \cos^{-1}\left(\frac{K(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{K(\mathbf{x}_i, \mathbf{x}_i) K(\mathbf{x}_j, \mathbf{x}_j)}}\right) \end{aligned} \tag{5.3}$$

where $\phi(\cdot)$ is a nonlinear mapping function and $K(\cdot,\cdot)$ is the kernel function. The cosine angle distance in the kernel space can be constructed using only the kernel function without considering the direct knowledge of the mapping function $\phi(\cdot)$. The angle between two samples is small (cosine of angle is high) if these samples are close to each other and vice versa.

Clustering based diversity (CBD)

Clustering techniques evaluate the distribution of the samples in a feature space and group the similar samples into the same clusters. In [14], the standard k-means clustering [34] was used in the diversity step of binary SVM AL technique. The aim of using clustering in the diversity step is to consider the distribution of uncertain samples and select the cluster prototypes as they are more sparse in the feature space (i.e., distant one another). Since the samples within the same cluster are correlated and provide similar information, a representative sample is selected for each cluster. In [14], the sample that is closest to the corresponding cluster center (called medoid sample) is chosen as representative sample.

5.3.3 Proposed combination of uncertainty and diversity techniques generalized to multiclass problems

In this chapter, each uncertainty technique is combined with one of the (binary) diversity techniques presented in the previous section. In the uncertainty step, the *m* most uncertain samples are selected using either MCLU or BLU. In the diversity step, the most diverse $h < m$ samples are chosen based on either ABD or CBD generalized to the multiclass case. Here, four possible combinations are investigated: 1) MCLU with ABD (denoted by MCLU-ABD), 2) BLU with ABD (denoted by BLU-ABD), 3) MCLU with CBD (denoted by MCLU-CBD), and 4) BLU with CBD (denoted by BLU-CBD).

Combination of uncertainty with ABD for multiclass SVMs (MCLU-ABD and BLU-ABD)

In the binary AL algorithm presented in [13], the uncertainty and ABD criteria are combined based on a weighting parameter λ. On the basis of this combination, a new sample is included in the selected batch *X* according to the following optimization problem:

$$t = \arg\min_{i \in I/X} \left\{ \lambda \left| f(\mathbf{x}_i) \right| + (1-\lambda) \left[\max_{j \in X} \frac{K(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{K(\mathbf{x}_i, \mathbf{x}_i) K(\mathbf{x}_j, \mathbf{x}_j)}} \right] \right\} \tag{5.4}$$

where I denotes the indices of unlabeled examples whose distance to the classification hyperplane is less than one, I / X represents the index of unlabeled samples of *I* that are not contained in *X*, λ provides the tradeoff between uncertainty and diversity, and *t* denotes the index of the unlabeled sample that will be included in the batch. The cosine angle distance between each sample of I / X and the samples included in *X* is calculated and the maximum value is taken as the diversity value of the corresponding sample. Then, the sum of the uncertainty and diversity values weighted by λ is considered to define the combined value. The unlabeled sample $\mathbf{x}_t$ that minimizes such value is included in *X*. This process is repeated until the cardinality of *X* ($|X|$) is equal to *h*. This technique guarantees that the selected unlabeled samples in *X* are diverse regarding to their angles to all the others in the kernel space. Since the initial size of *X* is zero, the first sample included in *X* is always the most uncertain sample of I (i.e., closest to the hyperplane). We generalize this technique to multiclass architectures presenting the MCLU-ABD and BLU-ABD algorithms.

```
Algorithm 2: MCLU-ABD
Inputs:
$\lambda$ (weighting parameter that tune the tradeoff between uncertainty and diversity)
$m$ (number of samples selected on the basis of their uncertainty)
$h$ (batch size)
Output:
$X$ (set of unlabeled samples to be included in the training set)
1. Compute $c(\mathbf{x})$ for each sample $\mathbf{x} \in U$ .
2. Select the set of $m$ unlabeled samples with lower $c(\mathbf{x})$ value (most uncertain) $\{\mathbf{x}_1^{MCLU}, \mathbf{x}_2^{MCLU}, ..., \mathbf{x}_m^{MCLU}\}$.
3. Initialize $X$ to the empty set.
4. Include in $X$ the most uncertain sample (the one that has the lowest $c(\mathbf{x})$ value).
Repeat
    5. Compute the combination of uncertainty and diversity with the following equation formulated for the multiclass architecture:
       $t = \underset{i \in I/X}{\arg\min} \left\{ \lambda \left| c(\mathbf{x}_i) \right| + (1-\lambda) \left[ \max_{j \in X} \frac{K(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{K(\mathbf{x}_i, \mathbf{x}_i) K(\mathbf{x}_j, \mathbf{x}_j)}} \right] \right\}$    (5.5)
       where $I$ denotes the set of indices of $m$ most uncertain samples and $c(\mathbf{x})$ is calculated as explained in the MCLU subsection (with $c_{\min}(\mathbf{x})$ or $c_{diff}(\mathbf{x})$ strategy).
    6. Include the unlabeled sample $\mathbf{x}_t$ in $X$.
Until $|X| = h$
7. The supervisor $S$ adds the label to the set of samples $\{\mathbf{x}_1^{MCLU-ABD}, \mathbf{x}_2^{MCLU-ABD}, ..., \mathbf{x}_h^{MCLU-ABD}\} \in X$ and these samples are added to the current training set $T$.
```

It is worth noting that the main difference between (5.4) and (5.5) is that the uncertainty in (5.5) is evaluated considering the confidence function $c(\mathbf{x}_i)$ instead of the functional distance $f(\mathbf{x}_i)$ as in the binary case.

```
Algorithm 3: BLU-ABD
Inputs:
$\lambda$ (weighting parameter that tune the tradeoff between uncertainty and diversity)
$m$ (number of samples selected on the basis of their uncertainty)
$h$ (batch size)
$q$ (number of unlabeled samples selected for each binary SVM in the BLU technique)
$L$ (total class number)
Output:
$X$ (set of unlabeled samples to be included in the training set)
1. Select the $q$ most uncertain samples from each of the $L$ binary SVM included in the multiclass OAA architecture (totally $\rho = qL$ samples are obtained).
2. Remove the redundant samples and consider the set of $m \le \rho$ patterns $\{\mathbf{x}_1^{BLU}, \mathbf{x}_2^{BLU}, ..., \mathbf{x}_m^{BLU}\}$.
3. Compute $c(\mathbf{x})$ for the set of $m$ samples as follows: if one sample is selected by more than one binary SVM, $c(\mathbf{x})$ is calculated as explained in the MCLU subsection (with $c_{\min}(\mathbf{x})$ or $c_{diff}(\mathbf{x})$ strategy); otherwise $c(\mathbf{x})$ is assigned to the corresponding functional distance $f(\mathbf{x})$.
4. Initialize $X$ to the empty set.
5. Include in $X$ the most uncertain sample (the one that has the lowest $c(\mathbf{x})$ value).
Repeat
    6. Compute the combination of uncertainty and diversity with the equation (5.5).
    7. Include the unlabeled sample $\mathbf{x}_t$ in $X$.
Until $|X| = h$
8. The supervisor $S$ adds the label to the set of patterns $\{\mathbf{x}_1^{BLU-ABD}, \mathbf{x}_2^{BLU-ABD}, ..., \mathbf{x}_h^{BLU-ABD}\} \in X$ and these samples are added to the current training set.
```

Combination of uncertainty with CBD for multiclass SVMs (MCLU-CBD and BLU-CBD)

The uncertainty and CBD were combined for binary SVM AL in [14]. The uncertain samples are identified according to the MS strategy based on their distance to the hyperplane. Then, the standard *k*-means clustering is applied in the original feature space to the unlabeled samples whose distance to the hyperplane (computed in the kernel space) is less than one (i.e., those that lie in the margin) and the *k*=*h* clusters are obtained. The medoid sample of each cluster is added to *X* (i.e., $|X| = h$), labeled by the supervisor *S* and moved to the current training set. This algorithm evaluates the distribution of the uncertain samples within the margin and selects the representative of uncertain samples based on standard *k*-means clustering. We extend this technique to multiclass problems. Here we define the MCLU-CBD and BLU-CBD algorithms.

Algorithm 4: MCLU-CBD

```
Inputs:
$m$ (number of samples selected on the basis of their uncertainty)
$h$ (batch size)
Output:
$X$ (set of unlabeled samples to be included in the training set)
1. Compute $c(\mathbf{x})$ for each sample $\mathbf{x} \in U$ .
2. Select the set of $m$ unlabeled samples with lowest $c(\mathbf{x})$ (with $c_{\min}(\mathbf{x})$ or $c_{diff}(\mathbf{x})$ strategy) value (most uncertain) $\{\mathbf{x}_1^{MCLU}, \mathbf{x}_2^{MCLU}, ..., \mathbf{x}_m^{MCLU}\}$.
3. Apply the $k$-means clustering (diversity criterion) to the selected $m$ most uncertain samples with $k=h$.
4. Calculate the $h$ cluster medoid samples $\{\mathbf{x}_1^{MCLU-CBD}, \mathbf{x}_2^{MCLU-CBD}, ..., \mathbf{x}_h^{MCLU-CBD}\}$, one for each cluster.
5. Initialize $X$ to the empty set and include in $X$ the set of $h$ patterns $\{\mathbf{x}_1^{MCLU-CBD}, \mathbf{x}_2^{MCLU-CBD}, ..., \mathbf{x}_h^{MCLU-CBD}\} \in X$
6. The supervisor $S$ adds the label to the set of $h$ patterns $\{\mathbf{x}_1^{MCLU-CBD}, \mathbf{x}_2^{MCLU-CBD}, ..., \mathbf{x}_h^{MCLU-CBD}\} \in X$ and these samples are added to the current training set.
```

Algorithm 5: BLU-CBD

```
Inputs:
$m$ (number of samples selected on the basis of their uncertainty)
$h$ (batch size)
$q$ (number of unlabeled samples selected for each binary SVM in the BLU technique)
$L$ (total class number)
Output:
$X$ (set of unlabeled samples to be included in the training set)
1. Select the $q$ most uncertain samples from each of the $L$ binary SVMs included in the multi-class OAA architecture (totally $\rho = qL$ samples are obtained).
2. Remove the redundant samples and consider the set of $m \le \rho$ patterns $\{\mathbf{x}_1^{BLU}, \mathbf{x}_2^{BLU}, ..., \mathbf{x}_m^{BLU}\}$.
3. Compute $c(\mathbf{x})$ for the set of $m$ samples as follows: if one sample is selected by more than one binary SVM, $c(\mathbf{x})$ is calculated as explained in the MCLU subsection (with $c_{\min}(\mathbf{x})$ or $c_{diff}(\mathbf{x})$ strategy); otherwise $c(\mathbf{x})$ is assigned to the corresponding functional distance $f(\mathbf{x})$.
4. Apply the $k$-means clustering (diversity criterion) to the selected $m$ most uncertain samples ($k=h$).
5. Calculate the $h$ cluster medoid samples $\{\mathbf{x}_1^{BLU-CBD}, \mathbf{x}_2^{BLU-CBD}, ..., \mathbf{x}_h^{BLU-CBD}\}$, one for each cluster.
6. Initialize $X$ to the empty set and include in $X$ the set of $h$ patterns $\{\mathbf{x}_1^{BLU-CBD}, \mathbf{x}_2^{BLU-CBD}, ..., \mathbf{x}_h^{BLU-CBD}\} \in X$
7. The supervisor $S$ adds the label to the set of $h$ patterns $\{\mathbf{x}_1^{BLU-CBD}, \mathbf{x}_2^{BLU-CBD}, ..., \mathbf{x}_h^{BLU-CBD}\} \in X$ and these samples are added to the current training set.
```

5.4 Proposed novel query function

Clustering is an effective way to select the most diverse samples considering the distribution of uncertain samples in the diversity step of the query function. In the previous section we generalized the CBD technique presented in [14] to the multiclass case. However, some other limitations can compromise its application: 1) the standard *k*-means clustering is applied to the original feature space and not in the kernel space where the SVM separating hyperplane operates, and 2) the medoid sample of each cluster is selected in the diversity step as the corresponding cluster representative sample (even if "more informative" samples in that cluster could be selected).

To overcome these problems, we propose a novel query function that is based on the combination of a standard uncertainty criterion for multiclass problems and a novel Enhanced CBD (ECBD) technique. In the proposed query function, MCLU is used with the difference $c_{diff}(\mathbf{x})$ strategy in the uncertainty step to select the *m* most uncertain samples. The proposed ECBD technique, unlike the standard CBD, works in the kernel space by applying the kernel *k*-means clustering [35], [36] to the *m* samples obtained in the uncertainty step to select the $h < m$ most diverse patterns. The kernel *k*-means clustering iteratively divides the *m* samples into *k*=*h* clusters ($C_1, C_2, ...C_h$) in the kernel space. At the first iteration, initial clusters $C_1, C_2, ...C_h$ are constructed assigning initial cluster labels to each sample [35]. In next iterations, a pseudo centre is chosen as the cluster center (the cluster centers in the kernel space $\phi(\mu_1), \phi(\mu_2), ...\phi(\mu_h)$ can not be expressed explicitly). Then the distance of each sample from all cluster centers in the kernel space is computed and each sample is assigned to the nearest cluster. The Euclidean distance between $\phi(\mathbf{x}_i)$ and $\phi(\mu_v)$, $v = 1, 2, ..., h$, is calculated as [35], [36]:

$$\begin{aligned} D^2(\phi(\mathbf{x}_i), \phi(\mu_v)) &= \left\|\phi(\mathbf{x}_i) - \phi(\mu_v)\right\|^2 \\ &= \left\|\phi(\mathbf{x}_i) - \frac{1}{|C_v|}\sum_{j=1}^{m}\delta(\phi(\mathbf{x}_j), C_v)\phi(\mathbf{x}_j)\right\|^2 \\ &= K(\mathbf{x}_i, \mathbf{x}_i) - \frac{2}{|C_v|}\sum_{j=1}^{m}\delta(\phi(\mathbf{x}_j), C_v)K(\mathbf{x}_i, \mathbf{x}_j) + \\ &\frac{1}{|C_v|^2}\sum_{j=1}^{m}\sum_{l=1}^{m}\delta(\phi(\mathbf{x}_j), C_v)\delta(\phi(\mathbf{x}_l), C_v)K(\mathbf{x}_j, \mathbf{x}_l) \end{aligned} \tag{5.6}$$

where $\delta(\phi(\mathbf{x}_j), C_v)$ shows the indicator function. The $\delta(\phi(\mathbf{x}_j), C_v)$=1 only if $\mathbf{x}_j$ is assigned to C_v, otherwise $\delta(\phi(\mathbf{x}_j), C_v)$=0. The $|C_v|$ denotes the total number of samples in C_v and is calculated as $|C_v| = \sum_{j=1}^{m}\delta(\phi(\mathbf{x}_j), C_v)$. As mentioned before, $\phi(\cdot)$ is a nonlinear mapping function from the original feature space to a higher dimensional space and $K(\cdot,\cdot)$ is the kernel function. The kernel *k*-means algorithm can be summarized as follows [35]:

1. The initial value of $\delta(\phi(\mathbf{x}_i), C_v)$, $i = 1, 2, ..., m$, $v = 1, 2, ..., h$, is assigned and *h* initial clusters $\{C_1, C_2, ...C_h\}$ are obtained.

2. Then $\mathbf{x}_i$ is assigned to the closest cluster.

$$\delta(\phi(\mathbf{x}_i), C_v) = \begin{cases} 1 & \text{if } D^2(\phi(\mathbf{x}_i), \phi(\mu_v)) < D^2(\phi(\mathbf{x}_i), \phi(\mu_j)) \quad \forall \mathrm{j} \neq v \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

3. The sample that is closest to μ_v is selected as the pseudo centre η_v of C_v.

$$\eta_v = \arg\min_{\mathbf{x}_i \in C_v} D(\phi(\mathbf{x}_i), \phi(\mu_v)) \quad (5.8)$$

4. The algorithm is iterated until converge, which is achieved when samples do not change clusters anymore.

After $C_1, C_2, ...C_h$ are obtained, unlike in the standard CBD technique, the most informative (i.e., uncertain) sample is selected as the representative sample of each cluster. This sample is defined as

$$\mathbf{x}_v^{MCLU-ECBD} = \arg\min_{\phi(\mathbf{x}_i)\in C_v} \left\{ c_{diff}(\mathbf{x}_i^{MCLU}) \right\} \qquad v = 1, 2, ..., h \quad (5.9)$$

where $\mathbf{x}_v^{MCLU-ECBD}$ represents the *v*-th sample selected using the proposed query function MCLU-ECBD and is the most uncertain sample of the *v*-th cluster (i.e., the sample that has minimum $c_{diff}(\mathbf{x})$ in the *v*-th cluster). Totally *h* samples are selected, one for each cluster, using (5.9).

In order to better understand the difference in the selection of the representative sample of each cluster between the query function presented in [14] (which selects the medoid sample as cluster representative) and the proposed query function (which selects the most uncertain sample of each cluster), Fig. 5.3 presents a qualitative example. Note that, for simplicity, the example is presented for binary SVM in order to visualize the confidence value $c_{diff}(\mathbf{x})$ as the functional distance (MS is used instead of MCLU). The uncertain samples are firstly selected based on MS for both techniques, and then the diversity step is applied. The query function presented in [14] selects medoid sample of each cluster (reported in blue in the figure), which however is not in agreement with the idea to select the most uncertain sample in the cluster. On the contrary, the proposed query function considers the most uncertain sample of each cluster (reported in red in the figure). This is a small difference with respect to the algorithmic implementation but a relevant difference from a theoretical viewpoint and for possible implications on results.

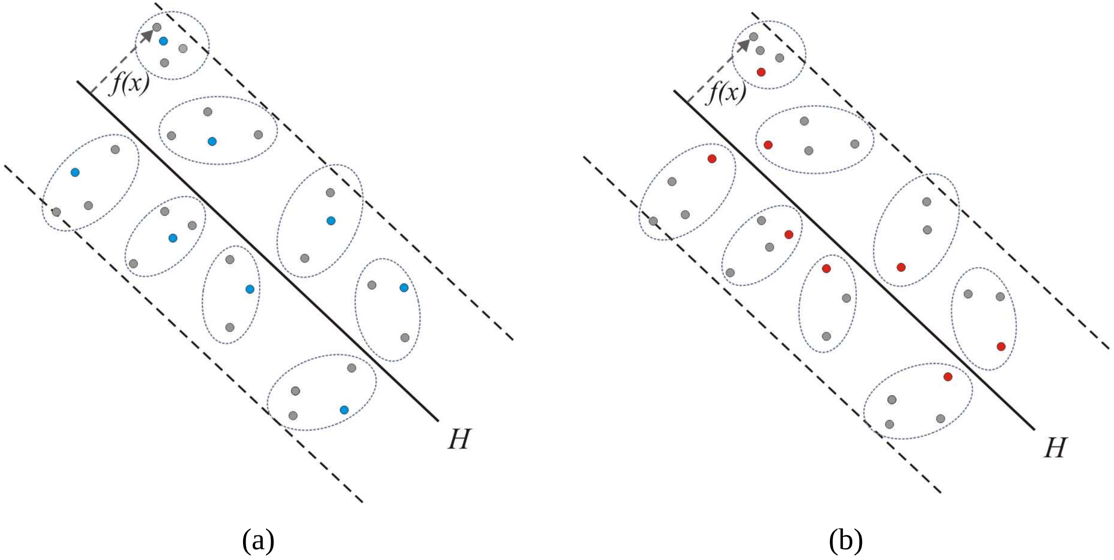


Fig. 5.3 - Comparison between the samples selected by (a) the CBD technique presented in [14], and (b) the proposed ECBD technique.

The proposed MCLU-ECBD algorithm can be summarized as follows:

```
Algorithm 6: Proposed MCLU-ECBD
Inputs:
$m$ (the number of samples selected on the basis of their uncertainty)
$h$ (batch size)
Output:
$X$ (set of unlabeled samples to be included in the training set)
1. Compute $c(\mathbf{x})$ for each sample $\mathbf{x} \in U$ .
2. Select the set of $m$ unlabeled samples with lower $c(\mathbf{x})$ value (most uncertain) $\{\mathbf{x}_1^{MCLU}, \mathbf{x}_2^{MCLU}, ..., \mathbf{x}_m^{MCLU}\}$.
3. Apply the kernel k-means clustering (diversity criterion) to the selected $m$ most uncertain samples with $k=h$.
4. Select the representative sample $\mathbf{x}_v^{MCLU-ECBD}, v=1,2,\ldots,h$ (i.e., the most uncertain sample) of each cluster according to (5.9).
5. Initialize $X$ to the empty set and include in $X$ the set of samples $\mathbf{x}_v^{MCLU-ECBD} \in X$ , $v=1,2,\ldots,h$.
6. The supervisor $S$ adds the label to the set of samples $\mathbf{x}_v^{MCLU-ECBD} \in X$ , $v=1,2,\ldots,h$, and these samples are added to the current training set.
```

5.5 Data set description and design of experiments

5.5.1 Data set description

Two data sets were used in the experiments. The first data set is a hyperspectral image acquired on a forest area on the Mount Bondone in the Italian Alps (near the city of Trento) on September 2007. This image consists of 1613×1048 pixels and 63 bands with a spatial resolution of 1 m. The available labeled data (4545 samples) were collected during a ground survey in summer 2007. The reader is referred to [37] for greater details on this dataset. The samples were randomly divided to derive a validation set *V* of 455 samples (which is used for model selection), a test set *TS* of 2272 samples (which is used for accuracy assessment), and a pool *P* of 1818 samples. The 4 % of the samples of each class are randomly chosen from *P* as initial training samples and the rest are considered as unlabeled samples. The land cover classes and the related number of samples used in the experiments are shown in Table 5.1.

The second data set is a Quickbird multispectral image acquired on the city of Pavia (northern Italy) on June 23, 2002. This image includes the four pan-sharpened multispectral bands and the panchromatic channel with a spatial resolution of 0.7 m. The image size is 1024×1024 pixels. The reader is referred to [1] for greater details on this dataset. The available labeled data (6784 samples) were collected by photointerpretation. These samples were randomly divided to derive a validation set *V* of 457 samples, a test set *TS* of 4502 samples and a pool *P* of 1825 samples. According to [1], Test pixels were collected on both homogeneous areas TS_1 and edge areas TS_2 of each class. The 4 % of the samples of each class in *P* are randomly selected as initial training samples, and the rest are considered as unlabeled samples. Table 5.2 shows the land cover classes and the related number of samples used in the experiments.

Table 5.1 - Number of samples of each class in *P*, *V* and *TS* for the Trento data set.

Class	***P***	**V**	***TS***
Fagus Sylvatica	720	180	900
Larix Decidua	172	43	215
Ostrya Carpinifolia	160	40	200
Pinus Nigra	186	47	232
Pinus Sylvestris	340	85	425
Quercus Pubescens	240	60	300
Total	1818	455	2272

Table 5.2 - Number of samples of each class in *P*, *V*, *TS1* and *TS2* for the Pavia data set.

Class	***P***	**V**	***TS*$_1$**	***TS*$_2$**
Water	58	14	154	61
Tree areas	111	28	273	118
Grass areas	103	26	206	115
Roads	316	79	402	211
Shadow	230	57	355	311
Red buildings	734	184	1040	580
Gray buildings	191	48	250	177
White building	82	21	144	105
Total	1825	457	2824	1678

5.5.2 Design of experiments

In our experiments, without loosing in generality, we adopt an SVM classifier with RBF kernel. The values for *C* and γ parameters are selected performing a grid-search model selection only at the first iteration of the AL process. Indeed, initial experiments revealed that, if a reasonable number of initial training samples is considered, performing the model selection at each iteration does not increase significantly the classification accuracies at the cost of a much higher computational burden. The MCLU step is implemented with different *m* values, defined on the basis of the value of *h* (i.e., $m = 4h,\ 6h, 10h$), with *h*=5,10,40,100. In the BLU technique, the *q*=*h* most uncertain samples are selected for each binary SVM. Thus the total number of selected samples for all SVMs is $\rho = qL$. After removing repetitive patterns, $m \leq \rho$ samples are obtained. The value of λ used in the MCLU-ABD and the BLU-ABD [for computing (5.5)] is varied as $\lambda = 0.3, 0.5, 0.6, 0.8$. The total cluster number *k* for both kernel *k*-means clustering and standard *k*-means clustering is fixed to *h*. All the investigated techniques and the proposed MCLU-ECBD technique are compared with the EQB and the MS-cSV techniques presented in [12]. The results of EQB are obtained fixing the number of EQB predictors to eight and selecting bootstrap samples containing 75 % of initial training patterns. These values have been suggested in [12]. Since the MS-cSV technique selects diverse uncertain samples according to their distance to the SVs, and can consider at most one sample related to each SV, it is not possible to define *h* greater than the total number of SVs. For this reason we can provide MS-cSV results for only *h*=5,10. Also the results obtained by the KL-Max technique proposed in [32] are provided for comparison purposes. Since the computational complexity of KL-Max implemented with SVM is very high, in our experiments at each iteration an unlabeled sample is chosen from a randomly selected subset (made up of 100 samples) of the unlabeled data. Note that the KL-Max technique can be implemented with any classifier that exploits posterior class probabilities for determining the decision boundaries [32]. In order to implement KL-Max technique with SVM, we converted the outputs of each binary SVM to posterior probabilities exploiting the Platt's method [39].

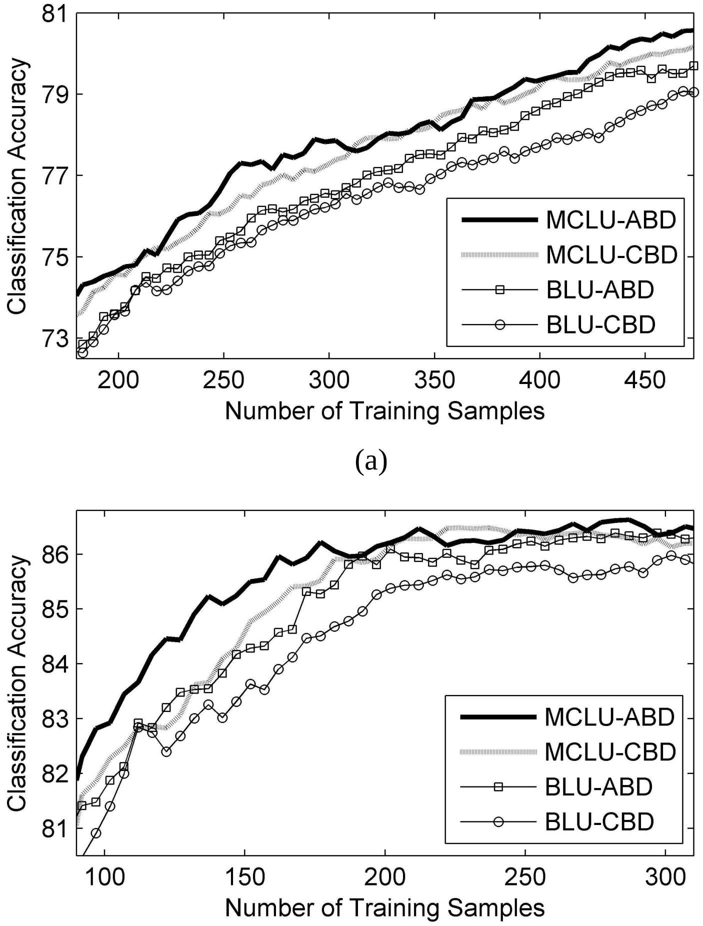
All experimental results are referred to the average accuracies obtained in ten trials according to ten initial randomly selected training sets. Results are provided as learning rate curves, which show the average classification accuracy versus the number of training samples used to train the SVM classifier. In all the experiments, the size of final training set $|T|$ is fixed to 473 for the Trento data set, and to 472 for the Pavia data set. The total number of iterations is given by the ratio between the number of samples to be added to the initial training set and the pre-defined value of *h*.

5.6 Experimental results

We carried out different kinds of experiments in order to: 1) compare the effectiveness of the different investigated techniques that we generalized to the multiclass case in different conditions; 2) assess the effectiveness of the novel ECBD technique; 3) compare the investigated methods and the proposed MCLU-ECBD technique with the techniques used in the RS literature; and 4) perform a sensitivity analysis with respect to different parameter settings and strategies.

5.6.1 Comparison among investigated techniques generalized to the multiclass case

In the first set of trials, we analyze the effectiveness of the investigated techniques generalized to multiclass problems. As an example, Fig. 5.4 compares the overall accuracies versus the number of initial training samples obtained by the MCLU-ABD, the MCLU-CBD, the BLU-ABD and the BLU-CBD techniques with $h = 5$, *k*=5 and $\lambda = 0.6$. In the MCLU, *m*=20 samples are selected for both data sets. In the BLU, $m \leq 30$ and $m \leq 40$ samples are chosen for the Trento and Pavia data sets, respectively. The confidence value is calculated with the $c_{diff}(\mathbf{x})$ strategy for both MCLU and BLU, as preliminary tests pointed out that by fixing the query function, the $c_{diff}(\mathbf{x})$ strategy is more effective than the $c_{\min}(\mathbf{x})$ strategy in case of using MCLU, whether it provides similar classification performance to the $c_{\min}(\mathbf{x})$ strategy when using BLU. Fig. 5.4 shows that the MCLU-ABD technique is the most effective on both the considered data sets. Note that similar behaviors are obtained by using different values of parameters (i.e., *m, h,* λ and *k*). The effectiveness of the MCLU and BLU techniques for uncertainty assessment can be analyzed by comparing the results obtained by combining them with the same diversity techniques under the same conditions (i.e., same values for parameters). From Fig. 5.4, one can observe that the MCLU technique is more effective than the BLU in the selection of the most uncertain samples on both data sets (i.e., the average accuracies provided by the MCLU-ABD are higher than those obtained by the BLU-ABD and a similar behavior is obtained with the CBD). This trend is confirmed by using different values of parameters (i.e., *m, h,* λ and *k*). The ABD and CBD techniques can be compared by combining them with the same uncertainty technique under the same conditions (i.e., same values for parameters). From Fig. 5.4, one can see that the ABD technique is more effective than the CBD technique. The same behavior can also be observed by varying the values of parameters (i.e., *m, h,* λ and *k*).



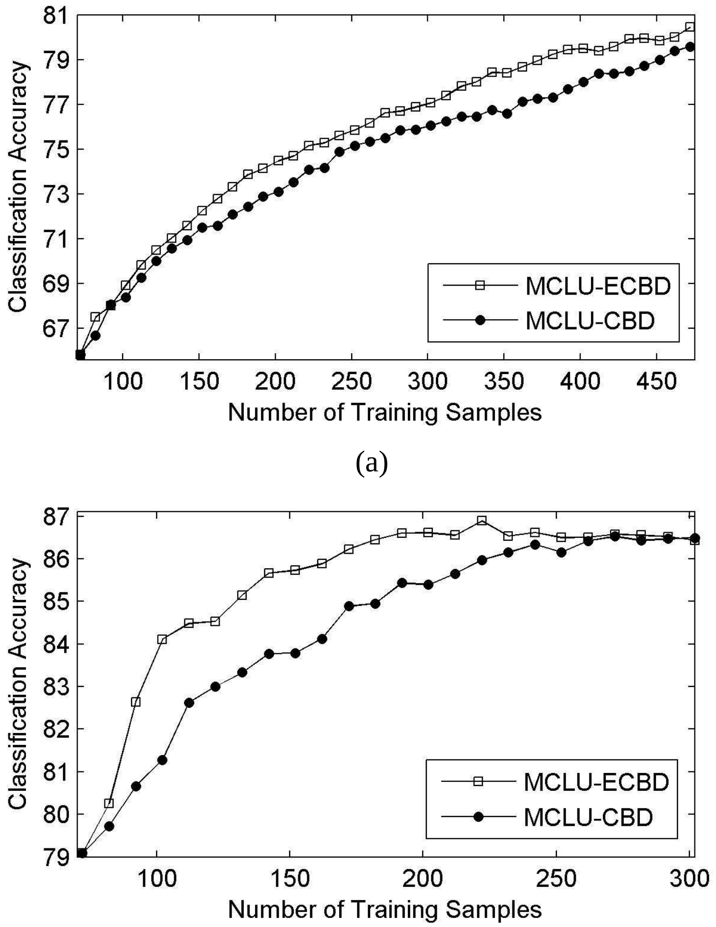
(a)

(b)

Fig. 5.4 - Overall classification accuracy obtained by the MCLU and BLU uncertainty criteria when combined with the ABD and CBD diversity techniques in the same conditions for (a) Trento, and (b) Pavia data sets. The learning curves are reported starting from 183 samples and 87 samples for Trento and Pavia data sets, respectively, in order to better highlight the small differences.

5.6.2 Results with the proposed MCLU-ECBD technique

In the second set of trials, we compare the standard CBD with the proposed ECBD using the MCLU uncertainty technique with the $c_{diff}(\mathbf{x})$ strategy and fixing the same parameter values. As an example, Fig. 5.5 shows the results obtained with $m = 40, h = 10, k = 10$ for both data sets. Table 5.3 (Trento data set) and Table 5.4 (Pavia data set) report the mean and standard deviation of classification accuracies obtained on ten trials versus different iteration numbers and different training data size $|T|$. From the reported results, one can see that ECBD technique provides the selection of more informative samples compared to CBD technique achieving higher accuracies than the standard CBD algorithm for the same number of samples. In addition, it can reach the convergence in less iterations. These results are also confirmed in other experiments with different values of parameters (not reported for space constraints).



(a)

(b)

Fig. 5.5 - Overall classification accuracy obtained by the MCLU uncertainty criterion when combined with the standard CBD and the proposed ECBD diversity techniques for (a) Trento, and (b) Pavia data sets.

Table 5.3 - Average classification accuracy (CA) and standard deviation (std) obtained on ten trials for different training data size $|T|$ and iteration numbers (Iter. Num) (Trento data set)

Technique	**$\|T\|$ = 163** (Iter.Num. 9)		**$\|T\|$ = 193** (Iter. Num. 12)		**$\|T\|$ = 333** (Iter. Num. 26)	
	CA	std	CA	std	CA	std
Proposed MCLU-ECBD	72.78	1.20	74.13	1.42	78.00	1.00
MCLU-CBD	71.55	1.57	72.88	1.62	76.47	1.10

Table 5.4 - Average classification accuracy (CA) and standard deviation (std) obtained on ten trials for different iteration numbers (Iter. Num) and training data size $|T|$ (Pavia data set)

Technique	**$\|T\|$ = 102** (Iter.Num. 3)		**$\|T\|$ = 142** (Iter. Num. 7)		**$\|T\|$ = 172** (Iter. Num. 10)	
	CA	std	CA	std	CA	std
Proposed MCLU-ECBD	84.10	1.66	85.66	1.29	86.23	1.09
MCLU-CBD	81.28	1.77	83.77	1.59	84.88	1.36

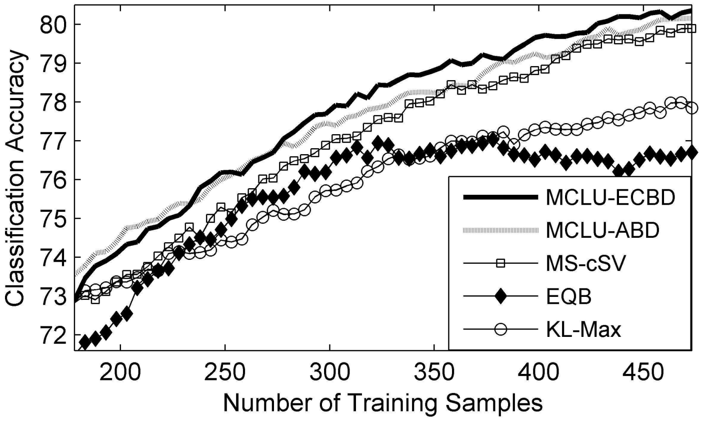
5.6.3 Comparison among the proposed AL techniques and literature methods

In the third set of trials, we compare the investigated and proposed techniques with AL techniques proposed in the RS literature. We compare the MCLU-ECBD and the MCLU-ABD techniques with the MS-cSV [31], the EQB [31] and the KL-Max [32] methods. According to the accuracies presented in section VA, we present the results obtained with the MCLU, which is more effective than the BLU. Fig. 5.6 shows the average accuracies versus the number of training samples obtained in the case of $h=5$ (*h*=1 only for KL-Max) for both data sets. For a fair comparison, the highest average accuracy result of each technique is given in the figure. Note that, since the MCLU-CBD proved less accurate than the MCLU-ECBD (see section V B), its results are no more reported here. For the Trento data set, the highest accuracies for MCLU-ECBD are obtained with $m=30$ (while *k*=5), whereas the best results for MCLU-ABD are obtained with λ=0.6 and $m=20$. For the Pavia data set, the highest accuracies for MCLU-ECBD are obtained with $m=20$ (while *k*=5), whereas the best results for MCLU-ABD are obtained with λ=0.6 and $m=20$.

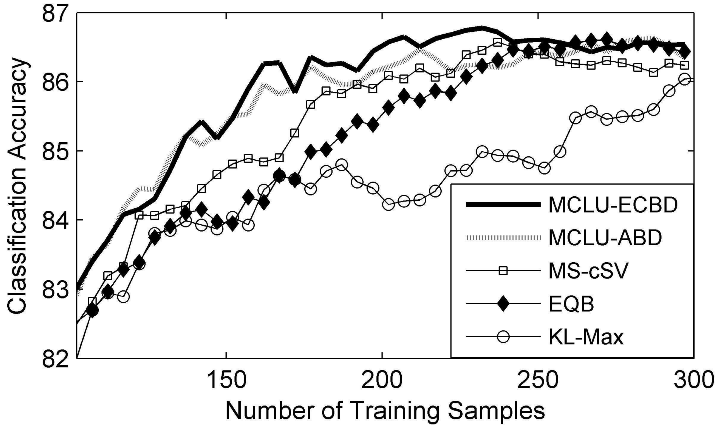
By analyzing Fig. 5.6(a) (Trento data set) one can observe that MCLU-ECBD and MCLU-ABD results are much better than MS-cSV, EQB, KL-Max results. The accuracy value at convergence of the EQB is significantly smaller than those of other techniques. The KL-Max accuracies are similar to the MS-cSV accuracies at early iterations. However, the accuracy of the KL-Max at convergence is smaller than those of the MCLU-ECBD and MCLU-ABD, as well as those of other methods. The results obtained on the Pavia data set [see Fig. 5.6(b)] show that the proposed MCLU-ECBD technique leads to the highest accuracies in most iteration; furthermore, it achieves convergence in less iterations than the other techniques. The MCLU-ABD method provides slightly lower accuracy than MCLU-ECBD; however, it results in significantly higher accuracies than MS-cSV, EQB as well as KL-Max techniques. KL-Max accuracy at convergence is significantly smaller than those achieved with other techniques.

For a better comparison, additional experiments were carried out on both data sets varying the values of the parameters. In all cases, we observed that MCLU-ECBD and MCLU-ABD yield higher classification accuracies than the other AL techniques when small *h* values are considered, and that the EQB technique is not effective when selecting a small number *h* of samples. On the contrary, the accuracies of EQB are close to those of MCLU-ECBD and MCLU-ABD when relatively high *h* values are considered. MS-cSV can not be used for high *h* values when small initial training set are available since the maximum number of *h* is equal to the total number of SVs. KL-Max results can only be provided for *h*=1 and the related accuracies are smaller than those of both MCLU-ECBD and MCLU-ABD methods.

Table 5.5 reports the computational time (in seconds) required by MCLU-ECBD, MCLU-ABD, MS-cSV, and EQB (for one trial) for different *h* values, and the computational time taken from KL-Max (related to *h*=1) for both data sets. In this case, the value of *m* for MCLU-ECBD and MCLU-ABD is fixed to 4*h* for both data sets. It can be noted that MCLU-ECBD and MCLU-ABD are fast both for small and high values of *h*. The computational time of MS-cSV and EQB is very high in the case of small *h* values, whereas it decreases by increasing the *h* value. The largest computational time is obtained with KL-Max that with an SVM classifier requires the use of the Platt algorithm for computing the class posterior probabilities. All the results clearly confirm that on the two considered data sets the proposed MCLU-ECBD is the most effective technique in terms of both computational complexity and classification accuracy.



(a)



(b)

Fig. 5.6 - Overall classification accuracy obtained by the MCLU-ECBD, MCLU-ABD, MS-cSV, EQB and KL-Max techniques for (a) Trento, and (b) Pavia data sets. The learning curves are reported starting from 178 samples and 92 samples for Trento and Pavia data sets, respectively, in order to better highlight the differences.

Table 5.5 - Examples of computational time (in seconds) taken from the MCLU-ECBD, MCLU-ABD, MS-cSV, EQB and KL-Max techniques

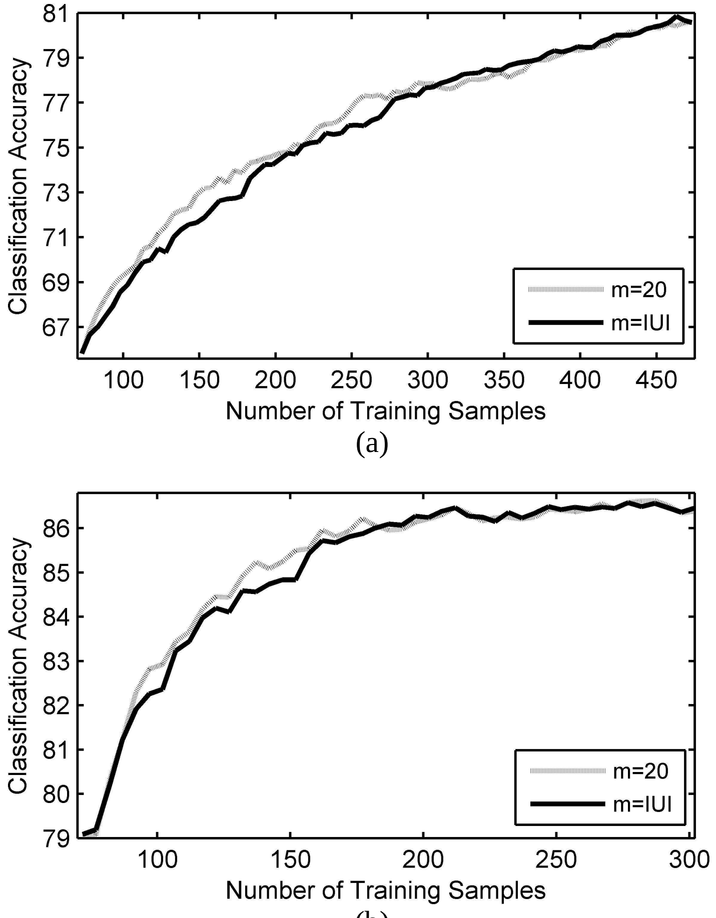
Data Set	**Technique**	***h***				
		1	**5**	**10**	**40**	**100**
Trento	**Proposed MCLU-ECBD**	-	10	6	8	12
	MCLU-ABD	-	10	6	7	10
	MS-cSV	-	584	452	-	-
	EQB	-	300	148	34	12
	KL-Max	72401	-	-	-	-
Pavia	**Proposed MCLU-ECBD**	-	10	6	7	11
	MCLU-ABD	-	10	5	6	10
	MS-cSV	-	384	193	-	-
	EQB	-	138	68	16	6
	KL-Max	71380	-	-	-	-

5.6.4 Sensitivity analysis with respect to different parameter settings and strategies

The aim of the fourth set of trials is to analyze the considered AL techniques under different parameter settings and strategies.

Analysis of the effect of the m value on the accuracy of the MCLU-ABD technique

We analyzed the effect of the *m* value on the classification accuracy obtained with the MCLU-ABD technique (which is the one that exhibited the highest accuracy among the investigated standard methods that we generalized to multiclass problems). In this technique, the equation (5.5) is calculated only for the *m* ($m = 4h,\ 6h, 10h$) most uncertain samples. The obtained results are compared to those obtained using all unlabeled samples, i.e., $m = |U|$. Fig. 5.7 shows the behavior of the overall classification accuracy versus the number of training samples obtained on both data sets with parameter values h=5, $m = 20$, $\lambda = 0.6$ and using the $c_{diff}(\mathbf{x})$ strategy. Results show that the choice $m = |U|$ produces accuracies close to those obtained using $m = 4h,\ 6h,\ 10h$ for both data sets. A similar behavior is observed in all the experiments carried out with different combinations of the abovementioned parameter values. Table 5.6 shows the computational time taken from the MCLU-ABD technique (for one trial) when $m = 4h$ and $m = |U|$, while h=5,10,40,100. From the table, one can observe that the value of m directly affects the computational time of MCLU-ABD: small m values decrease the computational time without resulting in a considerable loss in classification accuracy.



(a)

(b)

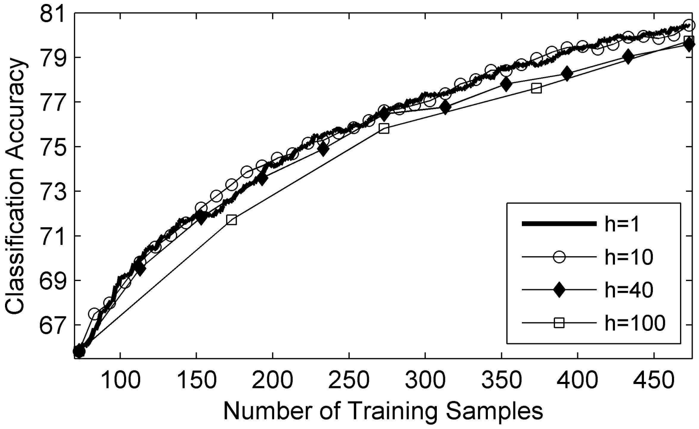
Fig. 5.7 - Overall classification accuracy versus the number of training samples obtained by the MCLU-ABD with respect to different m values for (a) Trento, and (b) Pavia data sets

Table 5.6 - Examples of computational time (in seconds) taken from the MCLU-ABD technique

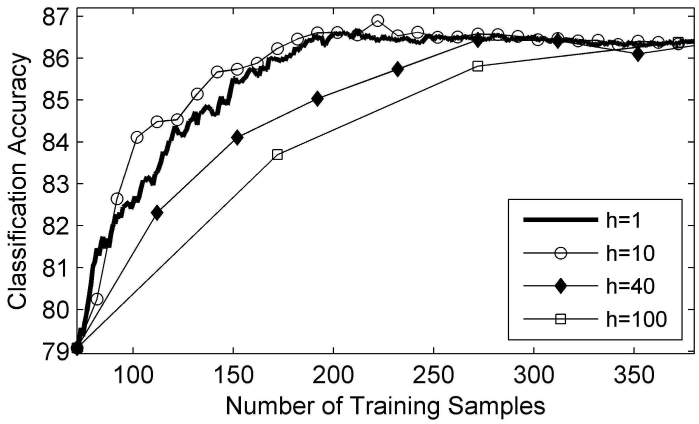
Data Set	***m***	***h***			
		5	**10**	**40**	**100**
Trento	$4h$	10	6	7	10
	$\lvert U \rvert$	37	36	35	35
Pavia	$4h$	10	5	6	10
	$\lvert U \rvert$	36	35	34	34

Analysis of the effect of different batch size values

We carried out an analysis of the performances of different AL techniques varying the value of the batch size h by fixing the query function. As an example, Fig. 5.8 shows the accuracies versus the number of training samples obtained on both data sets adopting the proposed MCLU-ECBD query function. The results are obtained with $m = 4h$ and $k = h$. The computational time taken from the MCLU-ECBD (related to one trial) for different *h* values is given in Table 5.7. From the table one can observe that the largest learning time is obtained in the case where one sample is selected at each iteration. The computational time decreases by increasing the *h* value. From Fig. 5.8, one can see that for both data sets selecting small *h* values results in similar (or better) classification accuracies compared to those obtained selecting only one sample at each iteration. On the contrary, high *h* values decrease the classification accuracy without decreasing the computational time if compared to small *h* values. Another interesting observation is that on the Pavia data set, when using small *h* values, convergence is achieved with less samples than when using large values. Note that similar behaviors are obtained with the other query functions.



(a)



(b)

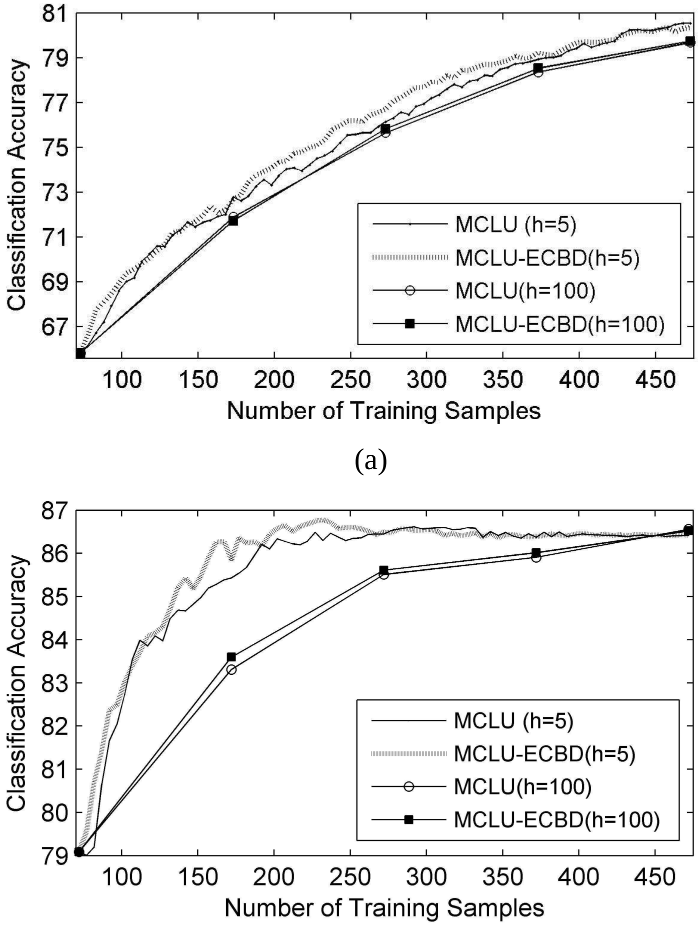
Fig. 5.8 - Overall classification accuracy versus the number of training samples obtained by the MCLU-ECBD technique with different *h* values for a) Trento and b) Pavia data sets

Table 5.7 - Examples of computational time (in seconds) taken from the MCLU-ECBD technique with respect to different *h* values

Data Set	**MCLU**	**MCLU-ECBD**		
	h	***h***		
	1	10	40	100
Trento	47	6	8	12
Pavia	46	6	7	11

Analysis of the effect of different batch size values h on the diversity criteria

Finally, we analyze the accuracy obtained by using only uncertainty criteria and the combination of uncertainty with diversity criteria for different *h* values. As an example, Fig. 5.9 shows the average accuracy versus the number of training samples obtained by MCLU (*m* is fixed to *h* for a fair comparison) and MCLU-ECBD with $m = 4h, h = 5{,}100$ and $k = h$. One can observe that, as expected, using only the uncertainty criterion provides poor accuracies when *h* is small, whereas the classification performances are significantly improved by using both uncertainty and diversity criteria. On the contrary, the choice of complex query functions is not justified when a large batch of samples is added to the training set at each iteration (i.e., similar results can be obtained with and without considering diversity). This mainly depends on the intrinsic capability of a large number of samples *h* to represent patterns in different positions of the feature space. Similar behaviors are observed with the other query functions.



(a)

(b)

Fig. 5.9 - Overall classification accuracy versus the number of training samples for the uncertainty criterion and the combination of uncertainty and diversity criteria with different *h* values: a) Trento and b) Pavia data sets

5.7 Discussion and conclusion

In this chapter, AL in RS classification problems has been addressed. Query functions based on MCLU and BLU in the uncertainty step, and ABD and CBD in the diversity step have been generalized to multiclass problems and experimentally compared on two different RS data sets. Furthermore, a novel MCLU-ECBD query function has been proposed. This query function is based on MCLU in the uncertainty step and on the analysis of the distribution of most uncertain samples by means of *k*-means clustering in the kernel space. Moreover, it selects the batch of samples at each iteration according to the identification of the most uncertain sample of each cluster.

In the experimental analysis we compared the investigated and proposed techniques with state-of-the-art methods adopted in RS applications for the classification of both a VHR multispectral and a hyperspectral image. By this comparison we observed that the proposed MCLU-ECBD method resulted in higher accuracy with respect to other state-of-the art methods on both the VHR and hyperspectral data sets. It was shown that the proposed query function is more effective than all the other considered techniques in terms of both computational complexity and classification accuracies for any *h* value. Thus, it is actually well-suited for applications which rely on both ground survey and image photointerpretation based labeling of unlabeled data. The MCLU-ABD method provides slightly lower accuracy than the MCLU-ECBD; however, it results in higher accuracies than the MS-cSV, the EQB as well as the KL-Max techniques. Moreover, we showed that: 1) the MCLU technique is more effective in the selection of the most uncertain samples for multiclass problems than the BLU technique; 2) the $c_{diff}(\mathbf{x})$ strategy is more precise than the $c_{\min}(\mathbf{x})$ strategy to assess the confidence value in the MCLU technique; 3) it is possible to have similar (sometimes better) classification accuracies with lower computational complexity when selecting small batches of *h* samples rather than selecting only one sample at each iteration; 4) the use of both uncertainty and diversity criteria is necessary when *h* is small, whereas high *h* values do not require the use of complex query functions; 5) the performance of the standard CBD technique can be significantly improved by adopting the ECBD technique, thanks to both the kernel *k*-means clustering and the selection of the most uncertain sample of each cluster instead of the medoid sample. In greater detail, on the basis of our experiments we can state that:

1) The proposed novel MCLU-ECBD technique shows excellent performance in terms of classification accuracy and computational complexity. It improves the already good performance of the standard CBD method. It is important to note that this technique has a computational complexity suitable to the selection of batch of samples made up of any desired number of patterns, thus it is compatible with both photointerpretation and ground survey based labeling of unlabeled data.

2) The MCLU-ABD technique provides slightly lower classification accuracies than the MCLU-ECBD method in most of the cases, with a similar computational time. It can be used for selecting a batch made up of any desired number of *h* samples. Thus, also the MCLU-ABD technique is suitable for both photointerpretation and ground survey based labeling of unlabeled data.

3) The MS-cSV technique provides quite good classification accuracies. However, the maximum value of *h* that can be used is equal to the total number of SVs $|\mathrm{SVs}|$ (i.e., $h \leq |\mathrm{SVs}|$ and therefore it can not be implemented for any *h* value). In the case of small *h* values, the computational complexity of this technique is much higher than that of the other investigated and proposed techniques. This complexity decreases when *h* increases. Therefore, the MS-cSV technique does not offer any advantage over the proposed technique.

4) The EQB technique results in poor classification accuracies with small values of *h* and classification accuracies comparable with other techniques with high values of *h*. The computational complexity of this technique is very high in case of selecting few samples, and decreases while *h* increases. Although it is possible to select any desired number of *h* samples with the EQB, it is not properly suitable for photointerpretation applications since its high computational complexity and poor classification performance with small *h* values. It is preferable for ground survey based labeling of unlabeled data.

5) The KL-Max technique is different from the above mentioned techniques since it is only able to select one sample at each iteration and can be implemented with any classifier that estimates a posteriori class probabilities. In our experiments we converted the SVM results into probabilities and results showed that this technique is not effective with SVM classifiers and requires very high computational complexity.

We assessed the compatibility of the considered AL techniques with the strategies to label unlabeled samples by image photointerpretation or ground data collection in order to provide some guidelines to the users under different conditions. As mentioned before, in the case of VHR images, in many applications the labeling of unlabeled samples can be achieved by photointerpretation, which is compatible with several iterations of the AL process in which a small value h of samples are included in the training set at each step according to an interactive procedure of labeling carried out by an operator. On our VHR data set, we observed that batches of $h = 5$ or 10 samples can give the highest accuracies. In the case of hyperspectral or medium/low resolution multispectral data, expensive and time consuming ground surveys are usually necessary for the labeling process. Under this last condition, only few iterations (two or three) of the AL process are realistic. Thus, large batches (of e.g., hundreds of samples) should be considered. In this case, we observed that sophisticated query functions are not necessary, as with many samples often an uncertainty criterion is sufficient for obtaining good accuracies. As a final remark, we point out that in real applications, some geographical areas may be not accessible for ground survey (or the process might be too expensive). Thus, the definition of the pool U should be carried out carefully, in order to avoid these areas. As a future development, we consider to extend the proposed method by including a spatially-dependent labeling costs, which takes into account that traveling to a certain area involves some type of costs (e.g., associated with gas or time) that should take into account in the selection of batch of unlabeled samples [27]. In addition, we plan to define hybrid approaches that integrate semisupervised and AL methods in the classification of RS images.

5.8 References

[1] B. M. Shahshahani and D. A. Landgrebe, “The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, vo. 5, pp. 1087–1095, September 1994.

[2] L. Bruzzone, M. Chi, M. Marconcini, “A Novel Transductive SVM for the Semisupervised Classification of Remote-Sensing Images”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 11, pp. 3363-3373, 2006.

[3] M. Chi, L. Bruzzone, “Semi-supervised Classification of Hyperspectral Images by SVMs Optimized in the Primal”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 6, Part 2, pp. 1870-1880, June 2007.

[4] M. Marconcini, G. Camps-Valls, L. Bruzzone, “A Composite Semisupervised SVM for Classification of Hyperspectral Images”, *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 2, pp. 234-238, 2009.

[5] G. Camps-Valls, T.V. Bandos Marsheva, and D. Zhou, “Semi-Supervised Graph-Based Hyperspectral Image Classification”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 10, pp. 3044-3054, October 2007.

[6] M. Li and I. Sethi, “Confidence-Based active learning”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1251-1261, 2006.

[7] D. D. Lewis and W. A. Gale, “A sequential algorithm for training text classifiers,” *in Proc. 17th Annu. International ACM-SIGIR Conf. Res. Dev. Inf. Retrieval*, W. B. Croft and C. J. van Rijsbergen, Eds., London, U.K., pp. 3–12, 1994.

[8] C. Campbell, N. Cristianini, and A. Smola, “Query Learning with Large Margin Classifiers”, *Proc. 17th International Conf. Machine Learning (ICML ’00)*, pp. 111-118, 2000.

[9] G. Schohn and D. Cohn, "Less is More: Active Learning with Support Vector Machines", *Proc. 17th Int'l Conf. Machine Learning (ICML '00)*, pp. 839-846, 2000.

[10] S. Tong and D. Koller, "Support Vector Machine Active Learning with Applications to Text Classification", *Proc. 17th International Conf. Machine Learning (ICML '00)*, pp. 999-1006, 2000.

[11] T. Luo, K. Kramer, D.B. Goldgof, L.O. Hall, S. Samson, A. Remsen, and T. Hopkins, "Active Learning to Recognize Multiple Types of Plankton," *J. Machine Learning Research*, vol. 6, pp. 589-613, 2005.

[12] P. Mitra, B. U. Shankar, and S. K. Pal, "Segmentation of multispectral remote sensing images using active support vector machines," *Pattern Recognition Letters*, vol. 25, no. 9, pp. 1067–1074, Jul. 2004.

[13] K. Brinker, "Incorporating Diversity in Active Learning with Support Vector Machines," *Proc. of the International Conference on Machine Learning*, Washington DC, pp. 59-66, 2003.

[14] Z. Xu, K. Yu, V. Tresp, X. Xu, and J. Wang, "Representative sampling for text classification using support vector machines," *25th European Conf. on Information Retrieval Research*, pp. 393-407, 2003.

[15] H. T. Nguyen and A. Smeulders, "Active learning using pre-clustering," *in Proc. 21th ICML*, Banff, AB, Canada, pp. 623-630, 2004.

[16] D. Cohn, Z. Ghahramani, and M.I. Jordan, "Active Learning with Statistical Models," *J. Artificial Intelligence Research*, vol. 4, pp. 129-145, 1996.

[17] K. Fukumizu, "Statistical Active Learning in Multilayer Perceptrons", *IEEE Transactions Neural Networks*, vol. 11 , no. 1, pp. 17-26, Jan. 2000.

[18] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction," *in Proc. ICML*, Williamstown, MA, 2001, pp. 441–448.

[19] H. S. Seung, M. Opper, and H. Smopolinsky, "Query by committee", *Proc. 5th Annu. ACM Workshop Comput. Learning Theory*, Pittsburgh, PA, pp. 287–294, 1992.

[20] Y. Freund, H.S. Seung, E. Shamir, and N. Tishby, "Selective Sampling Using the Query by Committee Algorithm," *Machine Learning*, vol. 28, pp. 133-168, 1997.

[21] I. Dagan and S. P. Engelson, "Committee-based sampling for training probabilistic classifiers," *in Proc. ICML*, San Francisco, CA, 1995, pp. 150–157.

[22] N. Abe and H. Mamitsuka, "Query learning strategies using boosting and bagging," *in Proc. ICML*, Madison, WI, pp. 1–9, 1998.

[23] V. N. Vapnik, The Nature of Statistical Learning Theory, 2nd ed., New York: Springer, 2001.

[24] F. Melgani, L. Bruzzone, "Classification of Hyperspectral Remote Sensing Images With Support Vector Machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778- 1790, Aug. 2004.

[25] S.C. Hoi, R. Jin, J. Zhu, and M.R. Lyu, "Batch Mode Active Learning and Its Application to Medical Image Classification," *Proc. 23rd Int'l Conf. Machine Learning (ICML '06)*, pp.417-424, June 2006.

[26] S.C. Hoi, R. Jin, J. Zhu, and M.R. Lyu, "Batch Mode Active Learning with Applications to Text Categorization and Image Retrieval", *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1233 – 1248, Sept. 2009.

[27] A. Liu, G. Jun, J. Ghosh, "Active learning of hyperspectral data with spatially dependent label acquisition costs", *IEEE Int. Geoscience and Remote Sensing Symposium 2009*, (IGARSS '09), Cape Town, South Africa, in press.

[28] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[29] Y. Zhang, X. Liao, and L. Carin, “Detection of Buried Targets Via Active Selection of Labeled Data: Application to Sensing Subsurface UXO”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 7, pp. 2535-2543, November 2004.

[30] Q. Liu, X. Liao, and L. Carin, “Detection of Unexploded Ordnance via Efficient Semisupervised and Active Learning”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 9, pp. 2558-2567, September 2008.

[31] D. Tuia, F. Ratle, F. Pacifici, M. Kanevski, and W. J. Emery, “Active Learning methods for remote sensing image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218 -2232, Jul. 2009.

[32] S. Rajan, J. Ghosh, and M. M. Crawford, “An active learning approach to hyperspectral data classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 4, pp. 1231-1242, Apr. 2008.

[33] A. Vlachos, “A stopping criterion for active learning,” *in Computer, Speech and Language*, vol. 22, no. 3, pp. 295-312, Jul. 2008.

[34] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Upper Saddle River, NJ: Prentice-Hall, 1988

[35] R. Zhang and A. I. Rudnicky, “A Large scale clustering scheme for kernel k-means,” *IEEE International Conference on Pattern Recognition*, 11-15 August 2002, Quebec, Canada, pp. 289-292.

[36] B. Scholkopf, A. Smola, and K. R. Muller, “Nonlinear component analysis as a kernel eigenvalue problem,” *Neural Computation*, vol.10, pp. 1299-1319, July 1998.

[37] M. Dalponte, L. Bruzzone, and D. Gianelle, “Fusion of hyperspectral and LIDAR remote sensing data for the estimation of tree stem diameters,” *IEEE Int. Geoscience and Remote Sensing Symposium 2009*, (IGARSS '09), Cape Town, South Africa, in press.

[38] L. Bruzzone, L. Carlin, “A Multilevel Context-Based System for Classification of Very High Spatial Resolution Images”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 9, pp 2587-2600, 2006.

[39] J.C. Platt, “Probabilities for SV Machines,” in Advances in Large Margin Classifiers, A. Smola, P. Bartlett, B.Schölkopf, and D. Schuurmans, Eds., MIT Press, pp. 61-74, 1999.

Chapter 6

6. A Novel Protocol for Accuracy Assessment in Classification of Very High Resolution Images

This chapter presents a novel protocol for the accuracy assessment of thematic maps obtained by the classification of very high resolution (VHR) images. As the thematic accuracy alone is not sufficient to adequately characterize the geometrical properties of high resolution classification maps, we propose a protocol that is based on the analysis of two families of indices: 1) the traditional thematic accuracy indices and 2) a set of novel geometric indices that model different geometric properties of the objects recognized in the map. In this context, we present a set of indices that characterize five different types of geometric errors in the classification map: 1) over-segmentation; 2) under-segmentation; 3) edge location; 4) shape distortion; and 5) fragmentation. Moreover, we propose a new approach for tuning the free parameters of supervised classifiers on the basis of a multiobjective criterion function that aims at selecting the parameter values that result in the classification map that jointly optimize thematic and geometric error indices. Experimental results obtained on Quickbird images show the effectiveness of the proposed protocol in selecting classification maps characterized by a better tradeoff between thematic and geometric accuracy than standard procedures based only on thematic accuracy measures. In addition, results obtained with Support Vector Machines (SVM) classifiers confirm the effectiveness of the proposed multiobjective technique for the selection of free parameter values for the classification algorithm.

6.1 Introduction

With the availability of very high resolution (VHR) images acquired by satellite multispectral scanners (e.g., GeoEye-1, Quickbird, Ikonos, SPOT 5), it is possible to acquire detailed information on the shape and the geometry of the objects present on the ground. This detailed information can be exploited by automatic classification systems to generate land-cover maps that ex-

This chapter is in press on the *IEEE Transactions on Geoscience and Remote Sensing*. Title: “A Novel Protocol for Accuracy Assessment in Classification of Very High Resolution Images”. Authors: C. Persello, L. Bruzzone.

hibit a high degree of geometrical details. The precision that the classification system can afford in the characterization of the geometrical properties of the objects present on the ground is particularly relevant in many practical applications, e.g., in urban area mapping, building characterization, target detection, crop fields classification in precision farming, etc.

In this context, it is necessary to further develop both algorithms for characterizing the textural and geometric information present in VHR images, and effective classification techniques capable to exploit these properties for increasing the classification accuracy. In the literature, several techniques have been proposed for the classification of VHR images. Among the others, we recall the use of texture, geometric features, and morphological transformations for characterizing the context of each single pixel, and the use of classification algorithms that can operate in large dimensional feature spaces (e.g., SVM) [1]-[5]. Nonetheless, a major open issue in classification of VHR images is the lack of adequate strategies for a precise evaluation of the quality of the produced thematic maps. The most common accuracy assessment methodology in classification of VHR images is based on the computation of thematic accuracy measures according to collected reference data. However, the thematic accuracy alone does not result sufficient for effectively characterizing the geometrical properties of the objects recognized in a map, because it assesses the correctness of the land-cover labels of sparse test pixels (or regions of interests) that do not model the actual shape of the objects in the scene. Thus, often maps derived by different classifiers (or with different parameter values for the same classifier) that have similar thematic accuracy exhibit significantly different geometric properties (and thus global quality). For this reason, in many real classification problems the quality of the maps obtained by the classification of VHR data is assessed also through a visual inspection. However, this procedure can provide just a subjective evaluation of the map quality that can not be quantified. Thus, it is important to develop accuracy assessment protocols for a precise, objective, and quantitative characterization of the quality of thematic maps in terms of both thematic and geometric properties [6]. These protocols could be used not only for assessing the quality of thematic maps generated by different classification systems, but also for better driving the model selection of a single classifier, i.e., the selection of the optimum values for the free parameter of a supervised categorization algorithm.

An important area in which some studies related to the aforementioned problem have been done in the past is that of landscape ecology. Some approaches have been proposed in the landscape ecology literature to compare different maps by considering the spatial structure of the landscape [7] (and thus not only the thematic accuracy). As an example, in [8] different comparison methods that consider both the spatial structure and the pixel-based overlap (i.e., the thematic accuracy) simultaneously are presented. However, these methods are developed in a different framework and do not consider the particular properties of classification maps derived form VHR remote sensing images and the issues related to the tuning of the free parameters of a classifier.

In this chapter we address the abovementioned problem by proposing a novel protocol for a precise, automatic, and objective characterization of the accuracy of thematic maps derived from VHR images. The proposed protocol is based on the evaluation of two families of indices: 1) thematic accuracy indices, and 2) a set of novel geometric indices that assess different properties of the objects recognized in the thematic map. The proposed protocol can be used to: 1) to objectively characterize the thematic and geometric properties of classification maps; 2) to select the map that better fit specific user requirements; or 3) to identify the map that exhibits in average best global properties if no specific requirements are defined. Moreover, we propose a novel approach for tuning the free parameters of supervised classification algorithms (e.g., SVM), which is based on the optimization of a multiobjective problem. The aim of this approach is to select the parameter values that result in a classification map that exhibits high geometric and thematic accuracies.

The chapter is organized into six sections. The next section presents the background on the assessment of thematic accuracy of land-cover maps. Section 6.3 describes the proposed accuracy assessment protocol, and discusses the two families of presented geometric and thematic indices. Section 6.4 illustrates the proposed multiobjective criterion for the tuning of the free parameters (model selection) of a classifier. Section 6.5 presents the obtained experimental results, while section 6.6 draws the conclusion of the chapter.

6.2 Background on thematic accuracy assessment of classification maps

In this section we briefly recall the main concepts on the procedures used to assess the thematic accuracy of a classification map obtained by a supervised classifier [9], [10]. In general, two main issues should be addressed: 1) the collection of the labeled samples for both training and testing a supervised algorithm (which may require the subdivision of the reference sample set in two or more disjoint sets) and 2) the choice of the statistical measure to evaluate the error (or accuracy) in pattern classification.

With respect to the first issue, several resampling methods have been proposed in the pattern recognition and statistical literature, e.g., resubstitution, holdout, leave-one-out, cross-validation, bootstrap [11]-[14]. Holdout is one of the most widely adopted resampling strategies in remote sensing applications. It consists in partitioning the available labeled samples in two independent sets or in directly collecting two independent sets of samples in separate areas of the scene. One set is used for training the classifier, the other one for assessing the classification accuracy. In some cases it is preferable to split the available samples in three sets: 1) one for training the algorithm (*training set*); 2) one for tuning the free parameters of the classifier (*validation set*); and 3) one for assessing the final accuracy (*test set*). Holdout is less computationally demanding with respect to other methods (e.g., leave-one-out and k-fold cross validation) and it is particularly reliable when the available labeled samples are acquired in two spatially disjoint portions of the scene. Indeed, in this case it is possible to asses the generalization capability of the classifier for test pixels that are spatially disjoint from the ones used for the training (which may present a different spectral behavior). With all the mentioned resampling methods, it is important to adopt a stratified approach, i.e., the training and test sets (or each of the *k* folds) should contain approximately the same proportions of the class labels as the original data set. Otherwise imbalanced and skewed results can be obtained.

With respect to statistical measures for accuracy evaluation, the complete description of the information that comes out from the comparison of the classification of test samples with the reference labeled data is given by the confusion (or error) matrix *E*. *E* is a square matrix of size $L \times L$ (where *L* is the number of information classes in the considered problem) defined as:

$$E = \begin{bmatrix} e_{11} & e_{12} & e_{13} & \dots & e_{1C} \\ e_{21} & e_{22} & \dots & \dots & \dots \\ e_{31} & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ e_{C1} & \dots & \dots & \dots & e_{CC} \end{bmatrix} \tag{6.1}$$

The generic element e_{ij} of the matrix denotes the number of samples classified into category *i* ($i = 1,...,L$) by the supervised classifier that are associated with label *j* ($j = 1,...,L$) in the reference data set. This representation is complete as the individual accuracy of each category is described along with both the errors of inclusion (commission errors) and errors of exclusion (omission errors) [9]. From the confusion matrix, different indices can be derived to summarize the information with a scalar value. Let us consider the sum of the elements of the row *i*, $e_{i+} = \sum_{j=1}^{L} e_{ij}$ (which is the number of samples classified into the category *i* in the classification map), and the sum of the elements of column *j*, $e_{+j} = \sum_{i=1}^{l} e_{ij}$ (which is the number of samples belonging to category *j* in the reference data set). Two commonly adopted indices are the overall accuracy (*OA*) and the kappa coefficient of accuracy (*kappa*), defined as:

$$OA = \frac{\sum_{i=1}^{L} e_{ii}}{e} \tag{6.2}$$

$$kappa = \frac{e\sum_{i=1}^{L} e_{ii} - \sum_{i=1}^{L} e_{i+}e_{+i}}{e^2 - \sum_{i=1}^{L} e_{i+}e_{+i}} \tag{6.3}$$

where *e* is the total number of test samples. *OA* represents the ratio between the number of samples that are correctly recognized by the classification algorithm with respect to the total number of test samples. The kappa coefficient of accuracy is a measure based on the difference between the actual agreement in the confusion matrix (as indicated by the main diagonal) and the chance agreement, which is indicated by the row and column totals (i.e., the marginals). The kappa coefficient is widely adopted as it uses also off-diagonal elements of the error matrix, and as it compensates for chance agreement. However, as pointed out in [15], kappa statistics has also unfavorable features. The main objection to the kappa coefficient is that it was introduced as a measure of agreement for two observers (see [16]). Thus, the kappa coefficient evaluates the departure from the assumption that two observers' ratings are statistically independent, rather than a measure of classification accuracy. For this reason, in [15] it is suggested to use other measures instead of kappa statistic, e.g., the class-averaged accuracy defined as:

$$CA = \frac{1}{L} \frac{\sum_{j=1}^{L} e_{jj}}{e_{+j}}, \tag{6.4}$$

or an alternative coefficient based on Kullback-Leibler information. We refer the reader to [9]-[11] for further details on accuracy assessment procedures in remote sensing image classification.

It is important to point out that all the abovementioned thematic accuracy measures do not consider the geometrical quality of the map under assessment and the shape of the objects present in the scene, thus resulting in the impossibility to assess the correctness of the geometry of the objects recognized by the classification algorithm. This is reasonable to evaluate the quality of classification maps obtained by medium or low resolution images, where the geometry of the objects is difficult to characterize. On the contrary, for adequately assessing the quality of classification maps obtained by VHR images, it is important to define indices capable to evaluate the geometrical properties of the maps, and to use them together with more traditional thematic indices.

6.3 Proposed protocol for accuracy assessment in VHR images

In this section we present the proposed protocol for accuracy assessment that is based on the computation of both thematic and geometric indices. The proposed procedure for thematic accuracy assessment is a simple refinement of the more traditional procedures described in the previous section, which takes into account particular properties of the classification of VHR images. On the contrary, the introduction of geometric indices to characterize the properties of the objects present in VHR images is one of the main contributions of the chapter. Thematic and geometric indices are described in the following two subsections, respectively.

6.3.1 Thematic error indices

When VHR images are considered, we can clearly identify two different contributions to the overall thematic accuracy: 1) the accuracy obtained on homogeneous areas, where pixels are characterized by the spectral signature of only one class, and 2) the accuracy obtained on borders of the objects and details, where pixels are associated with a mixture of the spectral signatures of different classes. These two contributions model the attitude of a classifier to correctly classifying homogeneous regions and high frequency areas, allowing a more precise assessment of the quality of the classification map. The classification of mixed pixels is a difficult task with *crisp* classifiers, which should decide for the predominant class in the area associated with the pixel (*fuzzy* classifiers may be adopted in their place for considering the contributions of the different land-cover types to the spectral signature associated with each single pixel [17]). The proposed thematic accuracy assessment consists of the calculation of two separate indices: 1) thematic accuracy on homogeneous areas, 2) thematic accuracy on edge areas. This is accomplished extending the holdout strategy by defining two independent test sets: one on homogeneous areas (pixel "inside" objects), the other one on edge areas (pixels on the boundaries of objects). This results in the calculation of two independent confusion matrices. Any index derived from the confusion matrices (e.g., overall accuracy, kappa coefficient, etc.) may be adopted to calculate the accuracy on the two separate test sets. It is worth noting that different indices provide different information and can be used together (see the next section for a detailed discussion on the combined use of multiple indices for the tuning of the free parameters of a supervised classifier).

6.3.2 Geometric error indices

The geometric accuracy of a classification map is related to its precision in reproducing the correct geometry, the shapes, and the boundaries of the objects (e.g., buildings, streets, fields, etc.) present in the scene under investigation. In this chapter, in order to quantify the geometric accuracy of maps characterized by very high spatial resolution, we define a set of object-based indices (error measures) that evaluate different geometric properties of the objects represented in a thematic map with respect to a reference map. Some of these indices are partially inspired to the measures used in the accuracy assessment of segmentation maps, while others are imported from different domains of image processing. These indices are computed by using a reference map that defines the exact shape, structure and position of a set $\mathbf{O}=\{O_1,O_2,...,O_d\}$ of d objects (e.g., buildings) adequately distributed in the scene under investigation and with different properties (see the example in Fig. 6.1). Generally, given the high resolution of VHR images, the map of reference objects can be easily defined by photointerpretation (few objects are sufficient for a good characterization of the properties of the map). Please note that the labels of the classes of the reference objects are not required for the computation of the geometric accuracy indices. In this way the evaluation of the geometric properties of the objects recognized in the map can be separated from the assessment of the thematic accuracy. Moreover, we do not require having reference objects for all the classes considered in the classification problem, but only for the classes for which the geometric properties are important and the precise shape can be easily defined (e.g., buildings, fields, lakes, bridges, etc).

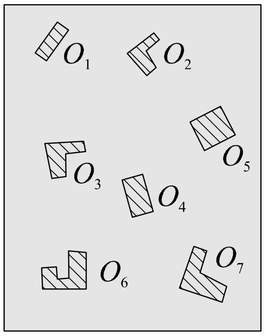


Fig. 6.1 – Example of a map of reference objects.

Let us consider that the thematic map under assessment (e.g., obtained by an automatic algorithm or by photointerpretation) is made up of a set $\mathbf{M}=\{M_1,M_2,...,M_r\}$ of r different regions of connected pixels (with 4- or 8-connectivity), such that each pixel in M_j, $j=1,2,...,r$, is associated with the same label v_j, where v_j is one of the L information classes in $\Omega=\{\omega_1,\omega_2,...,\omega_L\}$. In order to calculate the geometric error measures, it is necessary to identify for each object O_i in the reference map the corresponding region in the thematic map M_i. This can be done by considering the degree of overlapping between the pixels in the reference object O_i and in the regions M_j, $j=1,2,...,k$. The region M_i in the map with the highest overlapping area with the object O_i (i.e., with the highest number of common pixels) is selected according to:

$$M_i = \underset{\forall M_j \in \mathbf{M}}{\arg\max} \left| O_i \cap M_j \right| \qquad (6.5)$$

where $|\cdot|$ is the cardinality of a set, and here is used to extract the number of pixels (area) from a region (see the example in Fig. 6.2). Given a pair (O_i, M_i), it is possible to calculate a set of local geometric error measures $err_i^{(h)}$, $i=1,2,...,d$, $h=1,2,...,m$, that evaluate the degree of mismatching (in terms of m different specific geometric properties) between the reference object and the corresponding region on the map. Global error measures $err^{(h)}$, $h = 1, 2, ..., m$, can then be defined on the basis of the local measures.

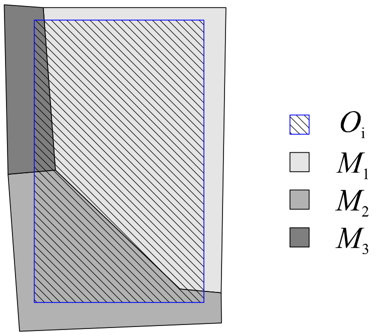


Fig. 6.2 – Example of a reference object O_i and the regions on the map that overlap with it. Region M_1 has the highest overlapping area with O_i and is selected according to (6.5).

The adopted measures are: 1) over-segmentation error, 2) under-segmentation error, 3) edge location error, 4) fragmentation error, and 5) shape error.

1) Over-segmentation - Similarly to the segmentation process, this error refers to the subdivision of a single object into several distinct regions in the classification map [see the example in Fig. 6.3(a)]. The proposed local error measure can be written as:

$$OS_i(O_i, M_i) = 1 - \frac{|O_i \cap M_i|}{|O_i|} \tag{6.6}$$

This measure evaluates the ratio between the overlapping area of the two regions (O_i, M_i) with respect to the area of the reference object. The index OS_i is defined in order to scale the output values in the range $[0,1)$. The higher is the value of the error, the higher is the level of over-segmentation of the object O_i in the considered classification map. The value of this error is 0 in the optimal case where the two regions are in full agreement, while it tends to 1 in the worst case of just one common pixel among the two regions.

2) Under-segmentation - The under-segmentation refers to the classification errors that result in group of pixels belonging to different objects fused into a single region [see the example in Fig. 6.3(b)]. The proposed local error measure is defined as:

$$US_i(O_i, M_i) = 1 - \frac{|O_i \cap M_i|}{|M_i|} \tag{6.7}$$

Unlike the over-segmentation, the under-segmentation error is computed by considering the ratio between the area of overlapping among M_i and O_i, and the area of the region on the map M_i. Also the US_i error varies in the range $[0,1)$. Value 0 of this index corresponds to perfect agreement between M_i and O_i, while values close to 1 reflect a high amount of under-segmentation (i.e., the region M_i is much bigger than the area of overlapping between the regions M_i and O_i).

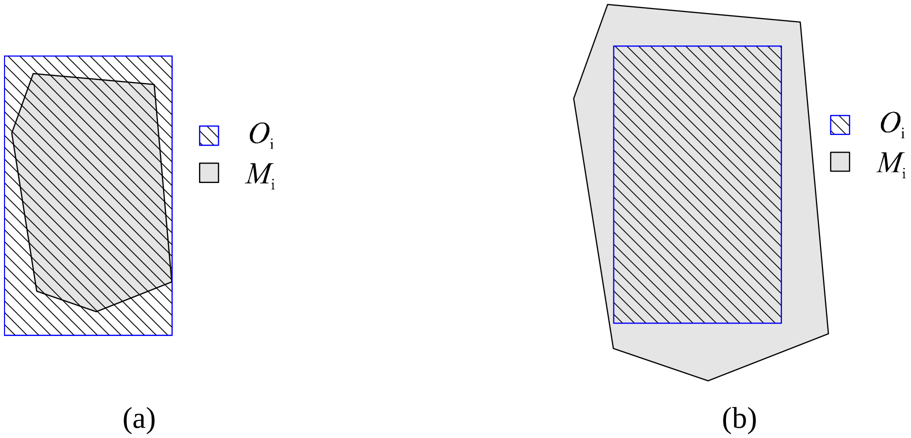


Fig. 6.3 – (a) Example of over-segmentation: the region M_i recognized on the map is smaller than the reference object O_i. (b) Example of under-segmentation: the region M_i recognized on the map is bigger than to the reference object O_i.

3) Edge location - This index measures the precision of the object edges recognized in the classification map with respect to those of the actual object [see the example in Fig. 6.4(a)]. Let $b(O_i)$ denote the operator that extracts the set of edge pixels from a generic region O_i. In this framework, we consider the possibility to introduce a tolerance in the recognition of the object borders. This can be implemented by adopting an operator $b(\cdot)$ that extracts the border line of the objects with a width greater than 1 pixel (e.g., 2 or 3 pixels). The definition of the border error is given by:

$$ED_i(O_i, M_i) = 1 - \frac{|b(O_i) \cap b(M_i)|}{|b(O_i)|} \tag{6.8}$$

This error measure varies in the range $[0,1)$ like the previous ones. A perfect matching in the borders of the two regions M_i and O_i leads to an error value equal to 0, whereas a large mismatching among the region edges results in error values close to 1.

4) Fragmentation error - The fragmentation of a classification map refers to the problem of sub-partitioning single objects into different small regions [see the example in Fig. 6.4(b)]. In order to quantitatively measure this type of error, we define a measure based on the number r_i of regions M_j, $j = 1, 2, ..., r_i$, that have at least one pixel in common with the reference object O_i. For this reason, we define the set $\mathbf{R}_i$ of all the regions overlapping with the reference object O_i as:

$$\mathbf{R}_i = \left\{ M_j, \forall j = 1, 2, ..., r_i : O_i \cap M_j \neq \varnothing \right\} \tag{6.9}$$

The proposed fragmentation error is then defined by the following equation:

$$FG_i(O_i, M_i) = \frac{r_i - 1}{|O_i| - 1} \tag{6.10}$$

This error value is scaled in a range $[0,1]$. The value is 0 in the optimal case when only one region M_j is overlapping with the reference object O_i, whereas it is 1 in the worst case where all the pixels of the object O_i belong to different regions M_j on the map. The measure is normalized with respect to the size (area) of the reference object O_i. It is worth noting that the fragmentation error is correlated with the over-segmentation error, but differs from the latter because it takes into account all the r_i regions M_j that overlap with the real object O_i, instead of the area of the single region M_i obtained by (6.5).

5) Shape error - This error is used to evaluate the shape difference between an object O_i and the corresponding region M_j on the map [see the example in Fig. 6.4(c)]. In order to characterize the shape of an object, several shape factors have been proposed in the literature and can be adopted (e.g., compactness, sphericity, eccentricity [18]). Thus the shape error can be defined as the absolute value of the difference in the selected shape factor $sf(\cdot)$ of the two regions M_i and O_i:

$$SH_i = \left\| sf(O_i) - sf(M_i) \right\| \tag{6.11}$$

It is worth noting that by adopting shape factors normalized in the range [0,1], the defined shape error measure will vary in the same range.

On the basis of the above defined measures of local errors (i.e., errors associated with single objects in the map), it is then possible to estimate global behaviors of the geometric properties of the classification map. Global error measurements can be obtained by averaging the local errors over the *d* measurements associated with the reference objects in $\mathbf{O}$, i.e., a generic global error measure characterizing property $err^{(h)}$ of the map can be expressed as:

$$err^{(h)} = \frac{1}{d}\sum_{i=1}^{d} err_i^{(h)} , \tag{6.12}$$

where $err_i^{(h)}$ is a local error *h* on the object *i*. In this way we give the same weight to the errors over the *d* objects, independently from their size. Other possible definitions of the global measures may take into account the size of the different objects, i.e.,

$$err^{(h)} = \frac{1}{d}\sum_{i=1}^{d} \left| O_i \right| err_i^{(h)} \tag{6.13}$$

or can weight differently the objects on the basis of specific user-defined requirements, i.e.,

$$err^{(h)} = \frac{1}{d}\sum_{i=1}^{d} \lambda_i \; err_i^{(h)} , \tag{6.14}$$

where λ_i, $i = 1, 2, ..., n$ are defined by the user. For example, the user may specify that geometric errors on buildings are more important than geometric errors on other objects, like streets, crop fields or lakes. Global measures are then used to estimate different geometric properties of the map. Combining the different global indices in a single measure that averages geometric indices is also possible. Nevertheless, this procedure would result in a measure that is difficult to understand.

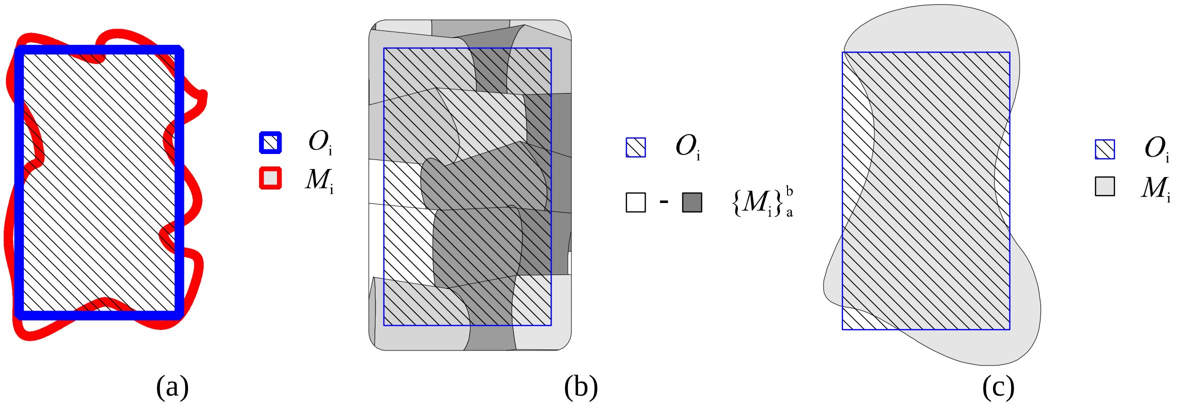


Fig. 6.4 – Example of a region M_i recognized on the map (with corresponding reference object O_i) that exhibits (a) large edge error, (b) high level of fragmentation, and (c) relatively high shape error.

6.4 Proposed multiobjective strategy for classifier parameter optimization

Other than the quality assessment of classification maps obtained according to different procedures (e.g., different automatic classifiers, photointerpretation, etc.), an accuracy index is also an important measure for tuning the free parameters of supervised classifiers (this process is also indicated as model selection). Let us consider a generic supervised algorithm for which a vector $\boldsymbol{\theta}$ of free parameters should be selected in order to optimize the quality of the output map. Standard approaches are based on the adoption of a scalar index to assess the thematic accuracy of the map (e.g., the overall accuracy or the kappa coefficient), and on the selection of the vector $\boldsymbol{\theta}$ that maximizes such a scalar value on the test samples. If a vector $\boldsymbol{I}$ of quality indices that characterize different thematic and geometric properties of the classification map is considered, the selection of $\boldsymbol{\theta}$ should be based on a different optimization strategy. The simplest (yet empirical and only partially reliable) strategy is to define a single error function $E(\cdot)$ combining the m proposed error measures according to a weighted average:

$$E(\mathbf{O},\mathbf{M}) = \sum_{j=1}^{m} c_j \; err^{(j)} \tag{6.15}$$

where the terms c_j, $j = 1, 2, ..., m$ are defined by the user. The set of parameter values of $\boldsymbol{\theta}$ that produces the classification map that minimizes $E(\mathbf{O},\mathbf{M})$ represents the solution to the considered problem. Nevertheless, this formulation has an important drawback: the definition of the c_j (which significantly affects the final result) is very critical because of the different intrinsic scales of the considered errors. In addition, the physical information conveyed by the resulting global index is difficult to understand.

To overcome this drawback, we propose to model our problem as a multiobjective minimization problem, where the multiobjective function $\mathbf{g}(\boldsymbol{\theta})$ is made up of m different objectives $g_1(\boldsymbol{\theta}), g_2(\boldsymbol{\theta}), ..., g_m(\boldsymbol{\theta})$ that represent the set of adopted error measures computed for different values of the classifier parameters (e.g., different thematic and geometric indices). All the different objectives of $\mathbf{g}(\boldsymbol{\theta})$ have to be jointly minimized and are considered equally important. In general all the proposed thematic indices (evaluated on homogeneous and border areas with different statistical parameters) and geometric indices could be used for the definition of $\mathbf{g}(\boldsymbol{\theta})$. However, depending on the application, it could be more appropriate to use different subsets of the presented indices as objectives of the optimization problem (e.g., for meeting some particular quality properties of the classification map required by the end users). Thus, the multiobjective problem can be formulated as follows:

$$\begin{gathered}\min_{\boldsymbol{\theta}\in S}\{\mathbf{g}(\boldsymbol{\theta})\},\;\; \mathbf{g}(\boldsymbol{\theta})=\left[g_1(\boldsymbol{\theta}),g_2(\boldsymbol{\theta}),...,g_m(\boldsymbol{\theta})\right]\\ \text{subject to } \boldsymbol{\theta}=(\theta_1,\theta_1,...,\theta_h)\in S\subseteq\mathbb{R}^h,\end{gathered} \tag{6.16}$$

where *S* denotes the search space for the classifier parameters. This problem is characterized by a vector-valued objective function $\mathbf{g}(\boldsymbol{\theta})$ and cannot be solved in order to derive a single solution like in optimization problems characterized by a single objective function. Instead, a set of optimal solutions P^* can be obtained by following the concept of Pareto dominance. In greater detail, a solution $\boldsymbol{\theta}^*$ is said to be Pareto optimal if it is not dominated by any other solution in the search space, i.e., there is no other $\boldsymbol{\theta}$ such that $g_i(\boldsymbol{\theta})\leq g_i(\boldsymbol{\theta}^*)$ $(\forall i=1,2,...,m)$ and $g_j(\boldsymbol{\theta})<g_j(\boldsymbol{\theta}^*)$ for at least one j ($j=1,2,...,m$). This means that $\boldsymbol{\theta}^*$ is Pareto optimal if there exists no other subset of classifier parameters $\boldsymbol{\theta}$ which would decrease an objective without simultaneously increasing another one (Fig. 3.2 clarifies this concept with a graphical example). The set P^* of all optimal solutions is called Pareto optimal set. The plot of the objective function of all solutions in the Pareto set is called Pareto front $PF^*=\{\mathbf{g}(\boldsymbol{\theta})\,|\,\boldsymbol{\theta}\in P^*\}$. The main advantage of the multiobjective approach is that it avoids to aggregate metrics capturing multiple objectives into a single measure. On the contrary, it allows one to effectively identify different possible tradeoffs between maps exhibiting different thematic and geometric properties.

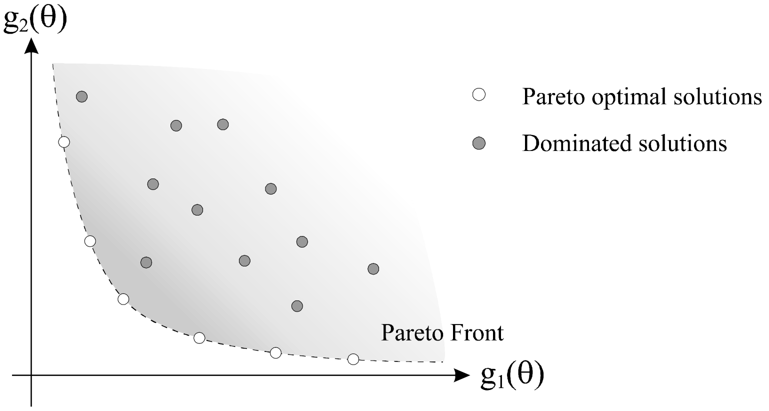


Fig. 6.5 – Example of Pareto-optimal solutions and dominated solutions in a two-objective search space.

Because of the complexity of the search space, an exhaustive search of the set of optimal solution P^* is unfeasible. Thus, instead of identifying the true set of optimal solutions, we aim to estimate a set of non-dominated solutions $\hat{P}^*$ with objective values as close as possible to the Pareto front. This estimation can be done with different multiobjective optimization algorithms [e.g., multiobjective evolutionary algorithms (MOEA) [19], [20]]. The final selection of the optimal solution among all estimated non-dominated solutions is demanded to the user, who can select the best tradeoff among the considered objectives on the basis of the specific application (e.g., one could tolerate to have under-segmented maps rather than over-segmented ones, or prefer to have less fragmented objects rather than high precision in the shape, etc.).

6.5 Experimental results

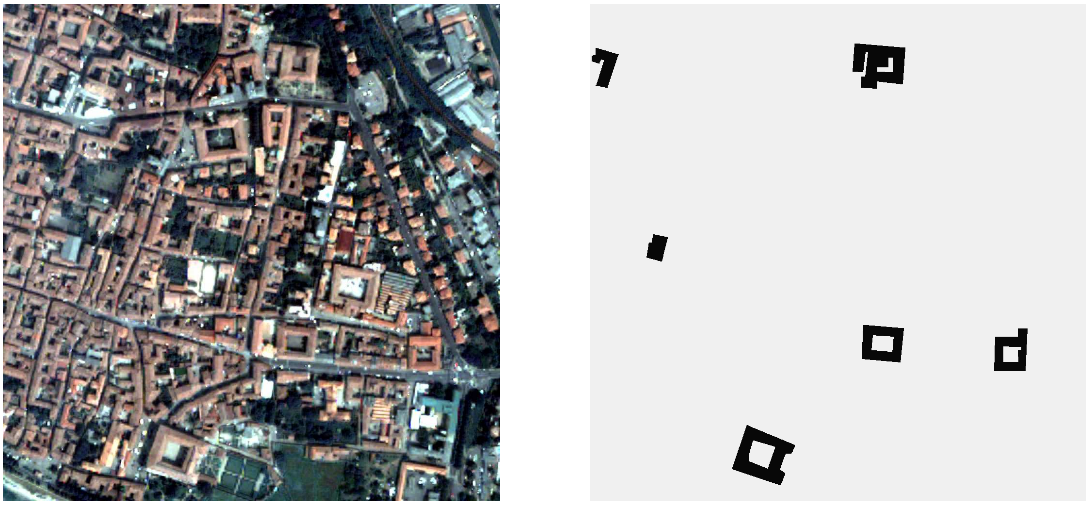
This section presents an experimental analysis aimed at studying the reliability of the proposed protocol for accuracy assessment of classification maps obtained by VHR images. We first applied the proposed indices to the quality assessment of different thematic maps obtained by the classification (carried out with different automatic techniques) of a Quickbird image acquired on the city of Pavia, Italy. Then, in a second set of experiments, we applied the proposed multiobjective strategy to the model selection of an SVM classifier in the analysis of a different Quickbird image acquired on the city of Trento, Italy. In our implementation of the geometric indices we considered a tolerance of 3 pixels for the edge location error, and we selected the eccentricity [18] as shape factor for the evaluation of the shaper error. The global geometric errors were computed on the basis of (6.12).

6.5.1 Quality assessment of classification maps

The first considered data set is made up of a Quickbird multispectral image acquired on the city of Pavia (northern Italy) on June 23, 2002. In particular, we used a panchromatic image and a pan-sharpened multispectral image [see Fig. 6.6(a)] obtained by applying a Gram Schmidt fusion technique [21] to the panchromatic channel and to the four bands of the multispectral image. The image size is 1024×1024 pixels with a spatial resolution of 0.7m. Greater details about this data set can be found in [1]. Table 6.1 presents the number of labeled reference samples for each set and class. Test pixels used for the assessment of thematic accuracy were collected on both edge and homogeneous areas. Test set pixels were taken from areas of the scene spatially disjoint from those related to the training samples. Fig. 6.6(b) shows the map of reference objects used for the evaluation of the geometric error indices. In particular, six different buildings were manually selected and considered as reference objects. It is worth noting that given the very high resolution of the images the procedure for digitizing few reference objects is simple and very fast.

Table 6.1 - Number of samples in the training and test sets (Pavia data set)

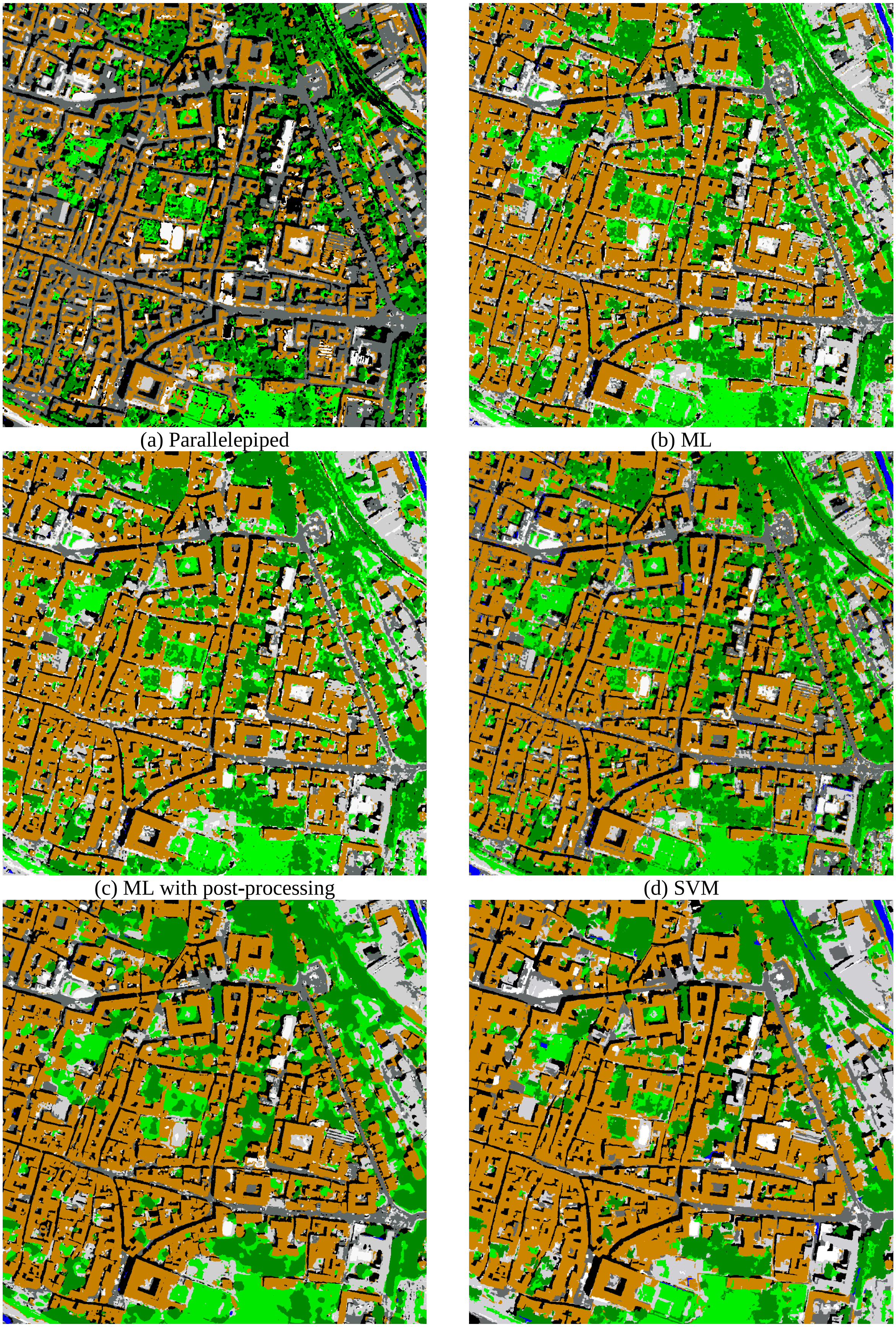
Class	**Number of patterns**		
	Training set	**Test set on edge areas**	**Test set on homogeneous areas**
Water	180	55	150
Tree areas	348	95	250
Grass areas	323	90	160
Roads	984	182	381
Shadow	750	297	325
Red buildings	2271	442	1040
Gray buildings	602	167	250
White building	275	98	100
TOTAL	5733	1426	2656



(a) (b)

Fig. 6.6 – (a) Real color composition of the image acquired by the Quickbird satellite on the city of Pavia (northern Italy). (b) Map of reference objects.

In our experiments, we obtained different thematic maps of the scene by using different automatic classification systems. The different systems were defined by varying the feature vector (i.e., considering only spectral features or also multiscale/multilevel contextual features), the supervised classification algorithms (i.e., parallelepiped, maximum likelihood, and SVM classifiers), and in some cases adding a post-processing phase for regularizing the final classification map. These systems were chosen with the goal to obtain classification maps with different properties. Fig. 6.7 shows the thematic maps obtained by the different considered classification systems. In particular, the maps (a)-(d) are obtained by considering a feature vector that is made up of only the original spectral features. Map (a) is obtained by using a very simple parallelepiped classifier (with $\sigma = 2$) [22]; map (b) is derived adopting a Gaussian Maximum Likelihood (ML) classifier; map (c) is obtained by applying a majority filter (with a sliding window of size 3×3) as post-processing to the map (b) [22]; map (d) is the result of the classification with SVM (using Gaussian kernels). The maps (e)-(h) are yielded using both spectral and contextual features, and adopting SVM as classification algorithm. Map (e) is obtained considering features extracted on the basis of the generalized Gaussian pyramid decomposition. In detail, the images were iteratively analyzed by a Gaussian kernel low-pass filter (with 5 × 5 square analysis window) and were under-sampled by factor two. We exploited five levels of pyramidal decomposition to characterize the spatial context of pixels and to label each pixel of the scene under investigation. Maps (f)-(h) are obtained using the multilevel context-based feature-extraction approach proposed in [1]; different statistical parameters are extracted from the pixels in each region defined at six different levels by a hierarchical segmentation process. In particular, for map (f) we considered the mean value for the first five levels and the standard deviation for the levels three, four, and five; for map (g) we considered only the mean for all first five levels. Map (h) is obtained considering the mean value extracted from all six segmentation levels.



(a) Parallelepiped

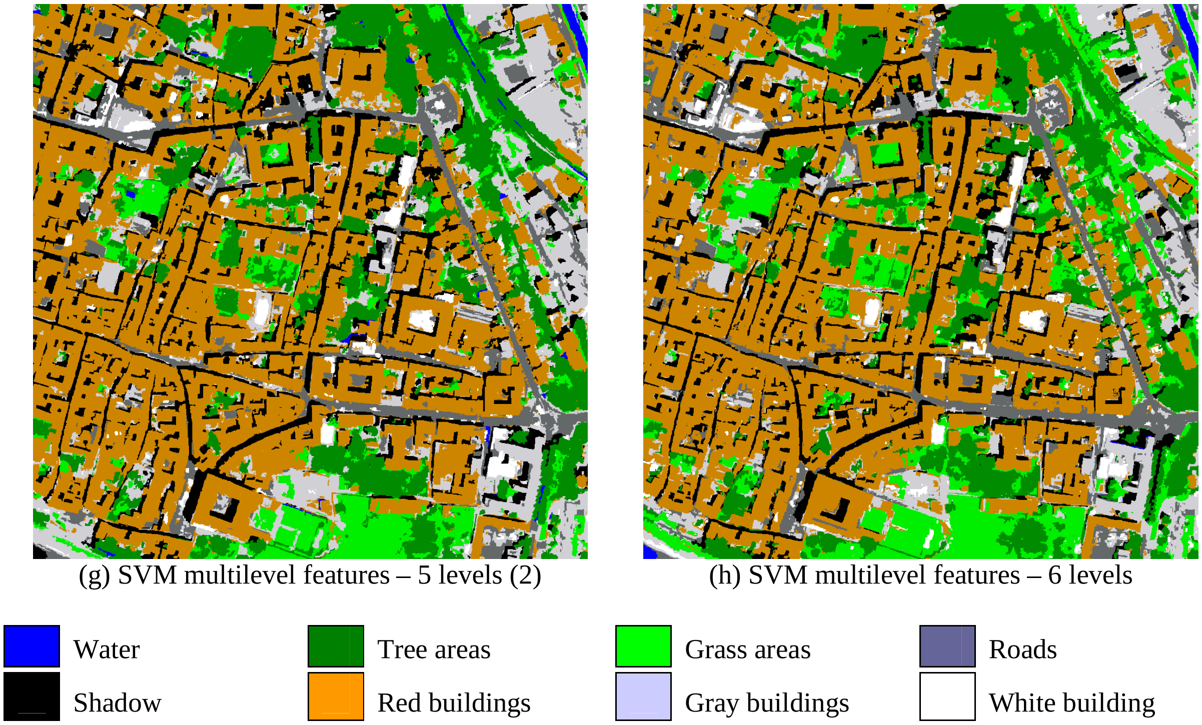
(b) ML

(c) ML with post-processing

(d) SVM

(e) SVM Gaussian Pyramid

(f) SVM multilevel features – 5 levels (1)


(g) SVM multilevel features – 5 levels (2)

(h) SVM multilevel features – 6 levels

Fig. 6.7 – Thematic maps obtained by different classification systems applied to the Pavia Quickbird image.

Table 6.2 and Table 6.3 report the thematic accuracies and the geometric error indices associated with the obtained maps, respectively. Considering the eight different maps, we can easily observe that, as expected, thematic maps obtained by pixel-based classification approaches [maps (a)-(b)-(d)] are less accurate than those obtained by context-based approaches. This general behavior is clearly pointed out also by thematic accuracy indices. The geometric error measurements give us important additional information about the different properties of the maps. In particular, we note that maps obtained by pixel-based approaches are generally more over-segmented and fragmented than the maps obtained by context-based classification systems, but they have also the important property to be less under-segmented. In the considered scene, we can observe that the buildings are very close each others. Thus, most of the considered classifiers merge regions associated to distinct objects (i.e., buildings) into a single region. The aforementioned problem is captured by the proposed geometric indices, which indicate that most of the maps have an under-segmentation error that is higher than the over-segmentation error [except for map (a)]. This problem strongly affects also the recognition of the correct shape of the objects. For this reason, we can observe that on this data set the shape error is highly correlated with the under-segmentation error. We can further observe that the edge location error is in general quite high for all obtained maps (even if a tolerance of 3 pixels is considered). This indicates that the considered classification techniques can scarcely model the correct borders of the objects.

Table 6.2 – Thematic accuracies computed on the test set on homogeneous areas, edge areas, and on both of them (complete test set) evaluated in terms of Overall Accuracy (*OA*) and kappa coefficient (*kappa*) (Pavia data set).

Map	**Complete Test set (homog. + edge areas)**		**Test set on homogeneous Areas**		**Test set on edge Areas**	
	OA%	***kappa***	***OA%***	***kappa***	***OA%***	***kappa***
(a) Parallelepiped	63.6%	0.564	77.7%	0.725	37.4%	0.287
(b) ML	83.0%	0.789	94.1%	0.925	62.4%	0.543
(c) ML post-processing	84.4%	0.805	95.2%	0.939	64.2%	0.564
(d) SVM	84.2%	0.801	94.7%	0.932	64.6%	0.563
(e) SVM Gaussian Pyramid	86.3%	0.828	95.0%	0.936	70.1%	0.631
(f) SVM Multilevel 5 levels (1)	90.0%	0.874	96.8%	0.960	77.1%	0.716
(g) SVM Multilevel 5 levels (2)	88.9%	0.861	97.1%	0.963	73.6%	0.677
(h) SVM Multilevel 6 levels	89.3%	0.866	96.2%	0.952	76.5%	0.711

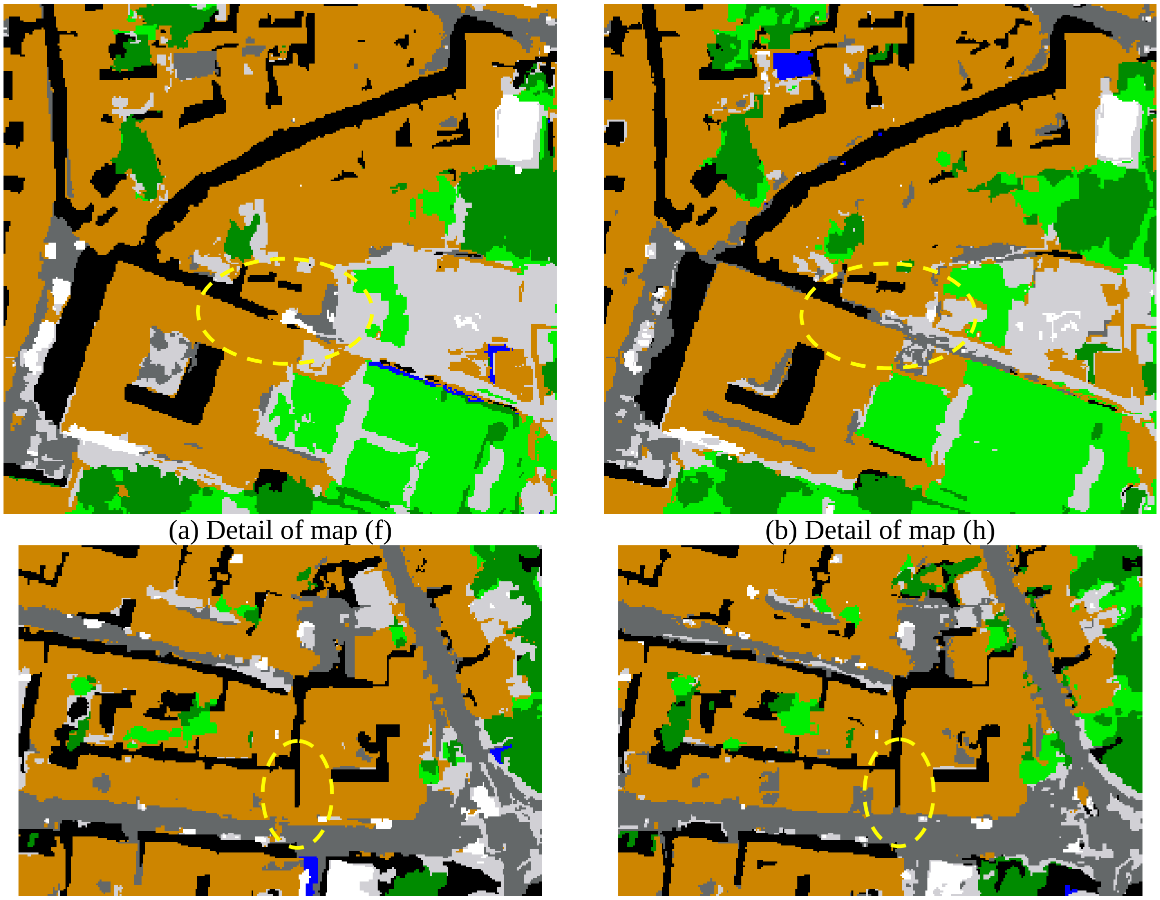
Table 6.3 – Geometric error indices (Pavia data set).

Map	**Under-segmentation**	**Over-segmentation**	**Edge location**	**Fragmentation**	**Shape**
(a) Parallelepiped	26.9 %	44.6 %	77.4 %	27.6 %	13.8 %
(b) ML	26.2 %	9.7 %	66.9 %	9.4 %	14.5 %
(c) ML post-processing	30.2 %	8.5 %	68.0 %	8.2 %	16.2 %
(d) SVM	16.9 %	12.7 %	58.4 %	7.4 %	12.9 %
(e) SVM Gaussian Pyramid	29.3 %	6.3 %	64.2 %	4.7 %	16.8 %
(f) SVM Multilevel 5 levels (1)	47.6 %	4.1 %	74.1 %	3.1 %	24.8 %
(g) SVM Multilevel 5 levels (2)	26.8 %	6.2 %	62.4 %	5.9 %	19.2 %
(h) SVM Multilevel 6 levels	27.1 %	4.8 %	58.5 %	4.3 %	17.0 %

Analyzing the single maps, we can observe that map (a) has very low quality in terms of thematic accuracy and in terms of most of the geometric indices. In particular, this map is sharply over-segmented and fragmented as indicated by the geometric errors; this is confirmed by a visual inspection. Map (b) has better quality than map (a): it exhibits higher thematic accuracy (both on homogeneous and border areas) and better geometric properties in terms of under-segmentation and border error. Map (c) [obtained by a post-processing applied to map (b)] results in slightly higher thematic accuracy, and in smaller over-segmentation, fragmentation and border errors than map (b). Nevertheless, the majority post-processing leads to slightly increase the under-segmentation error. Map (d) is the most accurate among those obtained with a pixel-based approach: this is pointed out by both thematic and geometric indices. In particular, this map exhibits the smallest under-segmentation and edge location errors among all considered maps. Map (e) exhibits important advantages with respect to the aforementioned maps, showing smaller over-segmentation and fragmentation errors as well as higher thematic accuracies. Nevertheless, the thematic accuracies (especially on border areas) are smaller than those of maps (f)-

(h). The geometric indices result particularly important for the characterization of the different maps obtained by the multilevel feature-extraction technique [maps (f)-(h)], which have high and very similar thematic accuracies. Map (f) is the most accurate from a thematic point of view, but maps (g) and (h) exhibit better geometric characteristics (e.g., under-segmentation and edge location error) than map (f). As it is possible to observe in Fig. 6.8, map (f) is affected by under-segmentation problems, as it merges different objects in the same region. On the contrary, map (h) correctly models the different buildings. This difference is clearly pointed out by the values of the under-segmentation error. Thus, considering both thematic and geometric indices, we can select map (h) as more reliable than map (f) (which would be preferred considering only thematic accuracies) because it presents a better tradeoff among different properties of the maps. It is worth noting that the property of correctly recognizing and distinguishing single objects in the scene can be very important for urban area analysis, especially in applications like building detection.

In general, the selection of the highest quality map depends on the kind of application and/or on end-user requirements. In this context, the proposed indices are a valuable tool that can drive the selection of the best thematic map in accordance to the application constraints.



(a) Detail of map (f)

(b) Detail of map (h)

(c) Detail of map (f)

(d) Detail of map (h)

Fig. 6.8 – Details of the thematic maps: (a)-(c) under-segmentation problems in map (f); (b)-(d) correct recognition of distinct buildings in map (h) (Pavia data set).

6.5.2 Multiobjective strategy for the model selection of supervised algorithms

In the second set of experiments, we used the proposed multiobjective technique for the model selection of a support vector machine (SVM) classifier with radial basis function (RBF) Gaussian kernels [23], [24]. The free parameters of the classifier are the regularization term C and the spread σ^2 of the Gaussian kernel. The experiments were carried out on a VHR image acquired by the Quickbird multispectral scanner on an urban area in the south of the city of Trento (Italy), on July 2006 [see Fig. 6.9(a)]. We used a panchromatic image and a pan-sharpened multispectral image obtained by applying a Gram Schmidt fusion technique to the panchromatic channel and to the four bands of the multispectral image. The image size is 500×500 pixels with a spatial resolution of 0.7 m. From the panchromatic and pan-sharpened multispectral bands we extracted textural features by applying an occurrence filter with 5×5 window size and computing mean, data range, and variance. Thus, the final feature vector is made up of 20 features (5 spectral features and 15 textural features). The available set of reference samples included a training set, a test set on homogeneous areas, and test set on border areas. The following six classes were considered: 1) roads, 2) red buildings, 3) dark buildings, 4) bright buildings, 5) shadow, and 6) vegetation. Table 6.4 presents the number of labeled reference samples for each set and class. Fig. 6.9(b) shows the map of the 11 reference objects used for the evaluation of the geometric error indices.

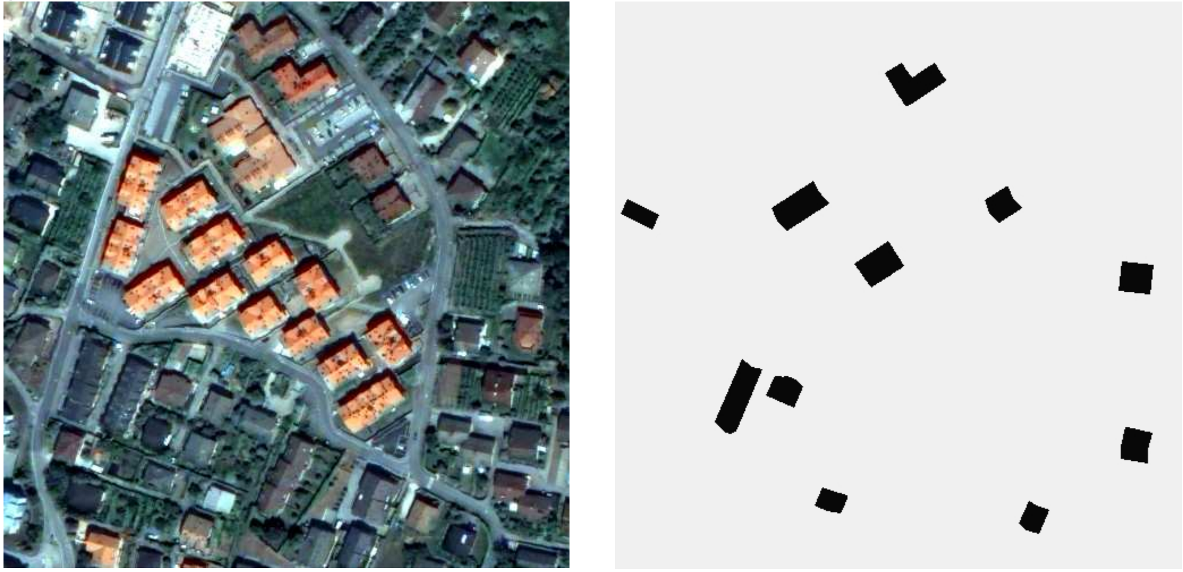


Fig. 6.9 – (a) Real color composition of the multispectral image acquired by the Quickbird satellite on the city of Trento (northern Italy). (b) Map of reference objects.

Table 6.4 - Number of samples in the training and test sets (Trento data set)

Class	Number of patterns		
	Training set	**Test set on edge areas**	**Test set on homogeneous areas**
Roads	58	63	47
Red roof buildings	71	70	73
Dark roof buildings	68	51	66
Bright roof buildings	39	38	39
Shadow	43	46	40
Vegetation	88	57	83
TOTAL	367	325	348

The strategy for the model selection proposed in section IV can be applied considering different sets of thematic and geometric indices as objectives of the optimization problem, depending on the specific application. In our analysis, we performed two sets of experiments considering: 1) seven objectives (two thematic and five geometric error indices), 2) two objectives (one thematic and one geometric error indices). These two sets of experiments represent examples of the practical use of the proposed multiobjective approach in real problems, but any other combination of thematic and geometric indices may be used in the optimization problem for the parameter tuning.

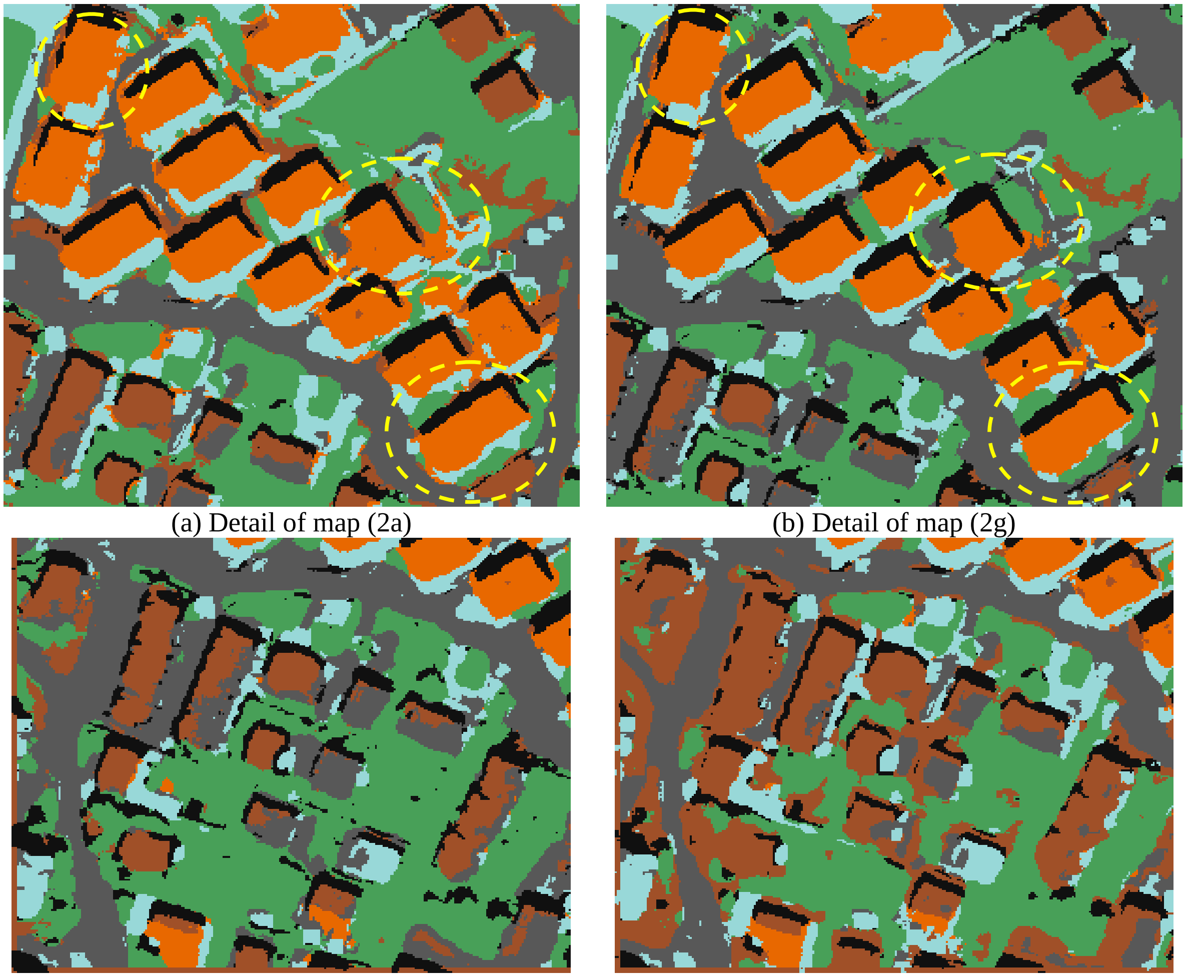
A) Experiments with seven error indices in the optimization problem

In this set of experiments we defined the model selection as a multiobjective optimization problem made up of seven objectives: the five geometric measures presented in Section III (i.e., under-segmentation, over-segmentation, edge location, fragmentation, and shape errors) and the two thematic errors based on kappa coefficient (calculated as 1-kappa) on the homogeneous and border test sets. Please note that in our experimental analysis we used a thematic error index based on the popular kappa coefficient, but any other index may be used in its place (e.g., the overall error). For the estimation of the Pareto-optimal solutions, we adopted a genetic multiobjective algorithm (a variation of NSGA-II) [25]. The population size was set to 30 and the maximum number of generation to 20. Among all Pareto-optimal solutions obtained by the genetic algorithm we selected seven solutions (used as an example in this discussion), characterized by different tradeoffs among the different indices (see Table 6.5). The selected solutions are characterized by the lowest error among all solutions for each index [e.g., map (2a) presents the highest thematic accuracy on homogeneous areas, map (2b) exhibits the highest thematic accuracy on edge areas, map (2c) exhibits the minimum under-segmentation error, etc.].

Table 6.5 – Thematic and geometric accuracy/error indices of seven solutions selected among all Pareto-optimal points estimated by the genetic algorithm. Each selected solution exhibit an accuracy index that has the highest value among all solutions (experiments with seven error indices in the optimization problem)

Map	**SVM parameters**		**Them. accuracies**		**Geometric errors**				
	C	$2\sigma^2$	**kappa homog.**	**kappa edge**	**Under-segment.**	**Over-segment.**	**Edge location**	**Fragm.**	**Shape**
(2a)	3187	0.820	0.951	0.892	12.7%	21.9%	51.3%	13.3%	14.7%
(2b)	4032	6.094	0.926	0.930	11.0%	25.5%	59.4%	15.4%	14.0%
(2c)	92	6.725	0.909	0.911	7.3%	29.9%	56.6%	13.1%	11.2%
(2d)	3464	0.221	0.926	0.771	31.9%	19.6%	63.6%	12.4%	16.8%
(2e)	2119	4.634	0.930	0.922	10.9%	25.2%	50.0%	15. 5%	13.9%
(2f)	4725	0.617	0.930	0.782	23.3%	21.4%	59.4%	10.5%	18.4%
(2g)	86	6.767	0.905	0.903	9.4%	30.2%	52.4%	12.7%	11.0%

All these Pareto-optimal solutions are associated with maps having different thematic and geometric properties. For example, Fig. 6.10 shows some details of the maps (2a)-(2g) and (2c)-(2d). Map (2a) (Fig. 6.10a) exhibits the highest kappa coefficient of accuracy on homogeneous areas, but the shape of red-roof buildings is not well recognized. On the contrary, map (2g) (Fig. 6.10b) has a smaller thematic accuracy, but better models the shape of the buildings. This behavior can also be observed by a visual inspection of the maps. Map (2c) (Fig. 6.10c) has the lowest under-segmentation error, whereas map (2d) (Fig. 6.10d) has good over-segmentation properties, in spite of significant under-segmentation errors (which also affect the recognition of the shape of the objects).



(a) Detail of map (2a)

(b) Detail of map (2g)

(c) Detail of map (2c)

(d) Detail of map (2d)

Fig. 6.10 – Details of maps associated with different Pareto-optimal solutions (experiments with seven error indices in the optimization problem, Trento data set).

B) Experiments with two error indices in the optimization problem

In this second set of experiments, two objectives were considered in the optimization problem: 1) the kappa coefficient of accuracy on homogeneous areas, and 2) the under-segmentation error. This represents an example in which we would like to select the SVM model that results in classification maps with the best tradeoff among thematic accuracy and precision in detecting separate buildings (under-segmentation error). The genetic algorithm adopted for the estimation of the Pareto front resulted in the estimation of the ten optimal solutions reported in Table 6.6.

Table 6.6 – Pareto-optimal solutions estimated by the genetic algorithm for the experiment with two error indices (Trento data set)

Map	**SVM parameters**		**Error indices**	
	C	$2\sigma^2$	**1-kappa (homog. areas)**	**Under-segmentation error**
(3a)	440	4.387	6.31%	9.95%
(3b)	155	5.836	6.66%	9.43%
(3c)	449	4.672	5.96%	10.36%
(3d)	799	1.177	5.61%	13.23%
(3e)	12	6.212	8.77%	6.86%
(3f)	24	6.455	7.01%	7.89%
(3g)	665	1.146	4.91%	13.62%
(3h)	667	1.145	5.26%	13.49%
(3i)	12	6.218	8.42%	7.19%
(3l)	19	6.059	7.71%	7.77%

Fig. 6.11 shows the estimated Pareto front. The selection of one model for the SVM classifier (i.e., the values of C and $2\sigma^2$) depends on the requirements of the specific application. For example, we selected three possible models from the Pareto-optimal solutions that leads to: 1) the map with the highest kappa coefficient of accuracy on homogeneous areas [map (3g)], 2) the map with the lowest under-segmentation error [map (3e)], 3) a good tradeoff between the two competing objectives [map (3c)]. A qualitative visual analysis of the obtained maps confirms that map (3g) [Fig. 6.12(a)] has some under-segmentation problems (but it has the smallest possible under-segmentation error for the obtained kappa value), map (3e) [Fig. 6.12(b)] is less under-segmented (and exhibits the highest possible kappa accuracy for the value of the obtained under-segmentation error), and map (3c) [Fig. 6.12(c)] can be considered a good tradeoff between the two considered objectives.

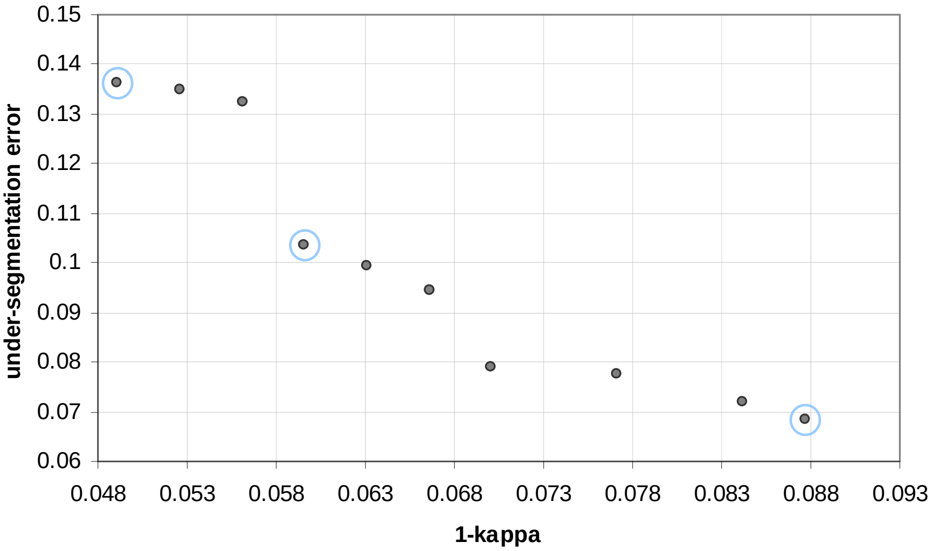
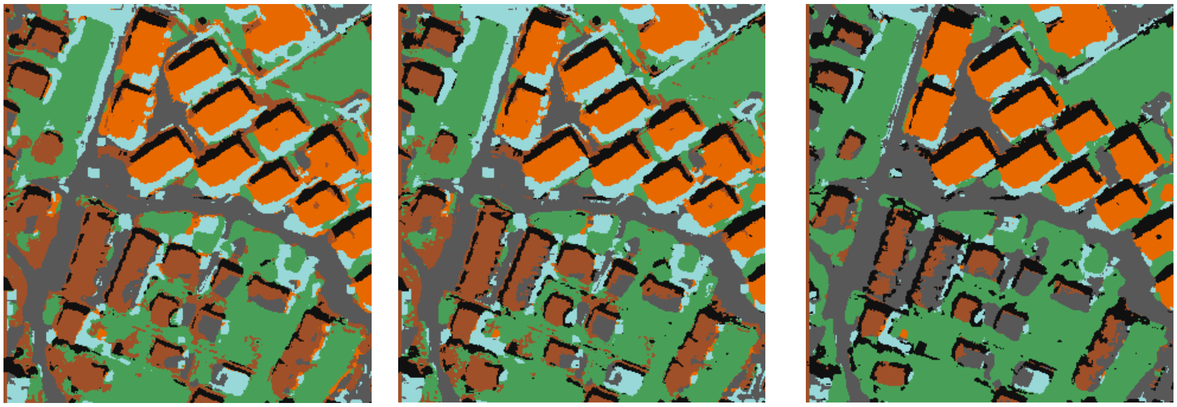


Fig. 6.11 – Estimated Pareto-optimal solutions for the experiment with two error indices in the optimization problem.



(a) Detail of map (3g) (b) Detail of map (3c) (c) Detail of map (3e)

Fig. 6.12 - Details of the maps associated with the three selected solutions (experiment with two error indices in the optimization problem, Trento data set).

It is worth noting that different error indices can be included in the multiobjective model selection. The choice of the error indices should reflect the properties that the end-users desire to optimize in the classification map. Other experiments, carried out using different error indices, confirmed the reliability of the proposed multiobjective model-selection technique based on the proposed accuracy assessment protocol.

6.6 Discussion and conclusion

In this chapter a novel protocol for the accuracy assessment of thematic maps obtained by the classification of VHR images has been presented. The proposed protocol is based on the evaluation of a set of error measures that can model the thematic and geometric properties of the obtained map. In particular, we presented a set of indices that characterize five different types of geometric errors in the classification map: 1) over-segmentation, 2) under-segmentation, 3) edge location, 4) shape distortion, and 5) fragmentation. The proposed geometric measures can be jointly used with the traditional thematic accuracy measures for a precise characterization of the properties of a thematic map derived by VHR images. The presented protocol can be used in three different frameworks: 1) assessing the quality of a classification map in an automatic, objective, and quantitative way; 2) selecting the classification map, among a set of different maps, that is more appropriate for the specific application on the basis of user-defined requirements; or 3) selecting the values of the free parameters of a supervised classification algorithm that result in the most appropriate classification map. Regarding this latter point, we have introduced a new technique for tuning the free parameters of supervised classifiers that is based on the optimization of a multiobjective problem, which results in parameter values that jointly optimize thematic and geometric error indices on the classification map.

Experimental results, obtained on two VHR images, confirms that the proposed geometric indices can accurately characterize the properties of classification maps, providing objective and quantitative error measures, which are in agreement with the observations derived by a visual inspection of the considered maps. Moreover, the proposed approach for tuning the free parameters of supervised classifiers resulted effective in the selection of the free parameters of SVM classifiers. This approach allows one to better characterize the tradeoff among the different thematic and geometric indices and to select the model in accordance with user requirements and application constraints.

It is worth noting that the proposed approach represents a step towards a new direction in accuracy assessment of classification maps derived from VHR images. However, some issues need to be further studied. One open issue is related to the definition of the reference objects and the evaluation of geometric indices in case of adjacent objects that can not be easily separated (e.g., different overlapped tree crown). Other issues are related to the definition of additional geometric indices to include in the proposed protocol and multiobjective strategy for considering different geometric properties. Moreover, as additional future developments of this research, we plan to extend the proposed multiobjective approach based on the evaluation of both thematic and geometric indices to the tuning of other variables of the classification system (not only related to the classification algorithm), e.g., for selecting the features to be given as input to the classifier or the parameters defining a post-processing, which strongly impact on the geometric properties of the final classification map.

6.7 References

[1] L. Bruzzone, L. Carlin, “A Multilevel Context-Based System for Classification of Very High Spatial Resolution Images”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 9, pp 2587-2600, 2006.

[2] J. A. Benediktsson, M. Pesaresi, and K. Arnason, “Classification and feature extraction for remote sensing images from urban areas based on morphological transformations,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 9, pp. 1940–1949, Sep. 2003.

[3] J. Chanussot, J. A. Benediktsson, and M. Fauvel, “Classification of Remote Sensing Images From Urban Areas Using a Fuzzy Possibilistic Model”, *IEEE Geoscience and Remote Sensing Letters*, vol. 3, no. 1, pp. 40-44, January 2006.

[4] P. Gamba, F. Dell'Acqua, G. Lisini, and G. Trinni, "Improved VHR Urban Area Mapping Exploiting Object Boundaries", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 8, pp. 2676-2682, August 2007.

[5] R. Bellens, S. Gautama, L. Martinez-Fonte, "Improved Classification of VHR Images of UrbanAreas Using Directional Morphological Profiles", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 10, pp. 2803-2813, October 2008.

[6] L. Bruzzone, C. Persello, "A Novel Protocol for Accuracy Assessment in Classification of Very High Resolution Multispectral and SAR Images", *Proc. IEEE International Geoscience and Remote Sensing Symposium*, (IGARSS '08), Boston, U.S.A., July 6 - 11, 2008, vol. 2, pp. II-265-II-268, 2008.

[7] R. White, "Pattern based comparisons", *Journal of Geographical Systems*, vol. 8, no. 2, pp. 145-164, May 2006.

[8] A. Hagen-Zanker, "Map coparison methods that simultaneously address overlap and structure", *Journal of Geographical Systems*, vol. 8, no. 2, pp. 165-185, May 2006.

[9] R. G. Congalton, K. Green, Assessing the Accuracy of Remotely Sensed Data, Lewis Publishers, Boca Raton, 1999.

[10] G. M. Foody, "Status of land cover classification accuracy assessment", *Remote Sensing of Environment*, no. 80, pp. 185-201, 2002.

[11] A. Baraldi, L. Bruzzone, P. Blonda, "Quality Assessment of Classification and Cluster Maps Without Ground Truth Knowledge," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 4, pp. 857-873, April 2005.

[12] R. Kohavi, "A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", *International Joint Conference on Artificial Intelligence* (IJCAI), 1995

[13] B. Efron, "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation", *Journal of the American Statistical Association*, vol. 78, no. 382, pp. 316-331, June 1983.

[14] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical Patter Recognition: A Review", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4-37, January 2000.

[15] R. Nishii, and S. Tanaka, "Accuracy and Inaccuracy Assessments in Land-Cover Classification", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 1, pp. 491-498, January 1999.

[16] J. Cohen, "A coefficient of agreement for nominal scales", *Education and Psychological Measurement*, vol. 20, no.1, pp. 37-40, 1960.

[17] F. Wang, "Fuzzy supervised classification of remote sensing images", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 28, no. 2, pp. 194–201, March 1990.

[18] K. R. Castleman, Digital Image Processing, Prentice Hall, Upper Saddle River, New Jersey, pp. 492-498, 1996

[19] C.M. Fonseca, P.J. Fleming, "Multiobjective optimization and multiple constraint handling with evolutionary algorithms-Part I: A unified formulation". *IEEE Transactions on Systems,. Man, and Cybernetics A*, vol. 28, no. 1, pp. 26-37, Jan 1998.

[20] A. Konak, D. W. Coit, A. E. Smith, "Multi-objective optimization using genetic algorithms: A tutorial", *Reliability Engineering and System Safety*, no. 91, pp. 992–1007, 2006.

[21] C. A. Laben and B. V. Brower, "Process for Enhancing the Spatial Resolution of Multispectral Imagery Using Pan-Sharpening," U.S. Pat. Office, US Patent 6,011,875, Washington, DC, 2000

[22] *ENVI user guide*, Available Online: http://www.ittvis.com/portals/0/pdfs/envi/Reference_Guide.pdf.

[23] M. Dalponte, M. Dalponte, L. Bruzzone, D. Gianelle, **"**Fusion of Hyperspectral and LIDAR Remote Sensing Data for Classification of Complex Forest Areas", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 5, pp. 1416 - 1427, May 2008.

[24] L. Bruzzone, C. Mingmin, M. Marconcini, “A Novel Transductive SVM for Semisupervised Classification of Remote-Sensing Images”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 11, pp. 3363 - 3373, Nov. 2006.

[25] K. Deb, Multi-objective optimization using evolutionary Algorithms, Chichester, John Wiley & Sons, 2001.

Chapter 7

7. Conclusions

This chapter concludes the thesis by summarizing and discussing the results obtained in the development of the considered research topics. Finally, it gives an outlook for future works.

7.1 Summary and discussion

In this thesis we investigated and developed different techniques and methods for the classification of VHR and hyperspectral RS images. In particular, we addressed several issues associated to different steps in the processing chain for the automatic classification of RS images (i.e., feature selection, classification techniques, accuracy assessment). We considered the following topics: 1) selection of a subset of the original features of a hyperspectral image that exhibits, at the same time, high capability to discriminate among the considered classes and high invariance in the spatial domain of the scene; 2) classification of RS images when the available training set is not fully reliable; 3) active learning techniques for interactive classification of RS images; and 4) definition of a protocol for accuracy assessment in the classification of VHR images that is based on the analysis of both thematic and geometric accuracy. For each considered topic detailed study of the literature was carried out and the limitations of currently published methodologies were highlighted. Starting from this analysis, novel solutions were theoretically developed, implemented and applied to real RS data in order to verify their effectiveness.

With respect to the first considered topic, in chapter 3 we have presented a novel feature-selection approach to the classification of hyperspectral images. The proposed approach aim at selecting subsets of features that exhibit, at the same time, high discrimination ability and high spatial invariance, improving the robustness and the generalization properties of the classification system with respect to standard techniques. The feature selection is accomplished by defining a multiobjective criterion function that considers the evaluation of both a standard separability measure and a novel term that measured the spatial invariance of the selected features. In order to assess the invariance in the scene of the feature subset, we proposed both a supervised method (assuming the availability of training samples acquired in two or more spatially disjoint areas) and a semisupervised method (which requires only a standard training set acquired in a single area of the scene and which exploits the information of unlabeled pixels in portions of the scene spatially disjoint from the training areas). The multiobjective problem is solved by an evolutionary algorithm for the estimation of the set of Pareto-optimal solutions. Experimental results showed that the proposed feature-selection approach selected subsets of the original features that sharply increased the classification accuracy on disjoint test samples, while it slightly decreased the accuracy on the adjoint test set with respect to standard methods. This behavior confirms that the proposed approach results in augmented generalization capability of the classification system. In particular, the proposed supervised method is effective in exploiting the information of the two available training sets, and the proposed semisupervised method can significantly increase the generalization capabilities of the classification system, without requiring additional reference data with respect to traditional feature-selection algorithms. This can be achieved at the cost of an acceptable additional computational time.

Concerning the second topic, in chapter 4 we have proposed a novel classification technique based on SVM that exploits the contextual information in order to render the learning of the classifier more robust to possible mislabeled patterns present in the training set. Moreover, we have analyzed the effects of mislabeled training samples on the classification accuracy of supervised algorithms, comparing the results obtained by the proposed CS^{4}VM with those yielded by a PS^{3}VM, a standard supervised SVM, a Gaussian ML, and a *k*-NN. This analysis was carried out varying both the percentage of mislabeled patterns and their distribution on the information classes. The experimental results obtained on two different data sets confirm that the proposed CS^{4}VM approach exhibits augmented robustness to noisy training sets with respect to all the other classifiers. In greater detail, the proposed CS^{4}VM method always increased the average kappa coefficient of accuracy of the binary classifiers included in the OAA multiclass architecture with respect to the standard SVM classifier. Moreover, in many cases, the CS^{4}VM sharply increased the accuracy on the information class that was most affected by the mislabeled patterns introduced in the training set. By analyzing the effects of the distribution of mislabeled patterns on the classes, it is possible to conclude that errors concentrated on a class (or on a subset of classes) are much more critical than errors uniformly distributed on all classes. In greater detail, when noisy patterns were added uniformly to all classes, we observed that the proposed CS^{4}VM resulted in higher and more stable accuracies than all the other classifiers. The supervised SVM and the PS^{3}VM exhibited relatively high accuracies when a moderate amount of noisy patterns was included in the training set, but they slowly decreased their accuracy when the percentage of mislabeled samples increased. On the contrary, both the ML and the *k*-NN classifiers are very sensitive even to the presence of a small amount of noisy patterns, and sharply decreased their accuracies by increasing the number of mislabeled samples. Nevertheless, the *k*-NN classifier resulted significantly more accurate than the ML classifier when mislabeled patterns equally affected the considered information classes. When noisy patterns were concentrated on a specific class of the training set, the accuracies of all the considered classifiers sharply decreased by increasing the amount of mislabeled training samples. Moreover, in this case, the proposed CS^{4}VM exhibited, in general, the highest and more stable accuracies. Nonetheless, when the number of mislabeled patterns increased over a given threshold, the classification problem became very critical and also the proposed technique significantly reduced its effectiveness. The standard SVM classifier still maintained higher accuracies than the ML and the *k*-NN techniques. The PS^{3}VM slightly increased the accuracies of the standard SVM. Unlike the previous case, the *k*-NN algorithm resulted in lower accuracies than the ML method. This is mainly due to the fact that mislabeled patterns concentrated on a single class (or on few classes) alter the prior probabilities, thus affecting more the *k*-NN classifier (which implicitly considers the prior probabilities in the decision rule) than the ML technique (which does not consider the prior probabilities of classes). The computational cost of the learning phase of the proposed CS^{4}VM method is slightly higher than that required from the standard supervised SVM. This depends on both the second step of the learning algorithm (which involves an increased number of samples, as semilabeled context patterns are considered in the process) and the setting of the additional parameters in the model-selection phase. However, the additional cost of the proposed method concerns only the learning phase, whereas the computational time in the classification phase remains unchanged.

In Chapter 5 we investigated the use of batch mode active learning for the interactive classification of RS images. Query functions based on MCLU and BLU in the uncertainty step, and ABD and CBD in the diversity step have been generalized to multiclass problems and experimentally compared on two different RS data sets. Furthermore, a novel MCLU-ECBD query function has been proposed. This query function is based on MCLU in the uncertainty step and on the analysis of the distribution of most uncertain samples by means of *k*-means clustering in the kernel space. Moreover, it selects the batch of samples at each iteration according to the identification of the most uncertain sample of each cluster. In the experimental analysis we compared the investigated and proposed techniques with state-of-the-art methods adopted in RS applications for the classification of both a VHR multispectral and a hyperspectral image. By this comparison we observed that the proposed MCLU-ECBD method resulted in higher accuracy with respect to other state-of-the art methods on both the VHR and hyperspectral data sets. It was shown that the proposed query function is more effective than all the other considered techniques in terms of both computational complexity and classification accuracies for any *h* value. Thus, it is actually well-suited for applications which rely on both ground survey and image photointerpretation based labeling of unlabeled data. The MCLU-ABD method provides slightly lower accuracy than the MCLU-ECBD; however, it results in higher accuracies than the MS-cSV, the EQB as well as the KL-Max techniques. Moreover, we showed that: 1) the MCLU technique is more effective in the selection of the most uncertain samples for multiclass problems than the BLU technique; 2) the $c_{diff}(\mathbf{x})$ strategy is more precise than the $c_{\min}(\mathbf{x})$ strategy to assess the confidence value in the MCLU technique; 3) it is possible to have similar (sometimes better) classification accuracies with lower computational complexity when selecting small batches of *h* samples rather than selecting only one sample at each iteration; 4) the use of both uncertainty and diversity criteria is necessary when *h* is small, whereas high *h* values do not require the use of complex query functions; 5) the performance of the standard CBD technique can be significantly improved by adopting the ECBD technique, thanks to both the kernel *k*-means clustering and the selection of the most uncertain sample of each cluster instead of the medoid sample.

In chapter 6 we have proposed a novel protocol for the accuracy assessment of thematic maps obtained by the classification of VHR images. The proposed protocol is based on the evaluation of a set of error measures that can model the thematic and geometric properties of the obtained map. In particular, we presented a set of indices that characterize five different types of geometric errors in the classification map: 1) over-segmentation, 2) under-segmentation, 3) edge location, 4) shape distortion, and 5) fragmentation. The proposed geometric measures can be jointly used with the traditional thematic accuracy measures for a precise characterization of the properties of a thematic map derived by VHR images. The presented protocol can be used in three different frameworks: 1) assessing the quality of a classification map in an automatic, objective, and quantitative way; 2) selecting the classification map, among a set of different maps, that is more appropriate for the specific application on the basis of user-defined requirements; or 3) selecting the values of the free parameters of a supervised classification algorithm that result in the most appropriate classification map. Regarding this latter point, we have introduced a new technique for tuning the free parameters of supervised classifiers that is based on the optimization of a multiobjective problem, which results in parameter values that jointly optimize thematic and geometric error indices on the classification map. Experimental results, obtained on two VHR images, confirms that the proposed geometric indices can accurately characterize the properties of classification maps, providing objective and quantitative error measures, which are in agreement with the observations derived by a visual inspection of the considered maps. Moreover, the proposed approach for tuning the free parameters of supervised classifiers resulted effective in the selection of the free parameters of SVM classifiers. This approach allows one to better characterize the tradeoff among the different thematic and geometric indices and to select the model in accordance with user requirements and application constraints.

7.2 Concluding remarks and future developments

In this research activity we developed techniques and approaches that can significantly improve the capability to automatically analyze and extract information from VHR and hyperspectral images. We addressed several issues related to feature selection for hyperspectral images, classification of VHR and hyperspectral data, and the definition of a novel protocol for the accuracy assessment of thematic maps obtained by the classification of VHR images. Moreover, we addressed operational problems related to the classification of RS images in real conditions where often the available reference samples are few and not completely reliable. The proposed methodologies contribute to a more effective use of last generation of RS data in many real-world applications related to the monitoring and the management of environmental resources.

Following the direction towards a more effective exploitation of last generation of RS images in real applications, several issues remain open and need to be addressed in future developments. Here, we identify (among the others) the following topics of interest: 1) feature selection/extraction in the kernel space for robust and accurate classification of hyperspectral images with kernel methods (e.g., support vector machine); 2) feature extraction methods for the classification of VHR images based on thematic and geometric accuracy indices; 3) classification techniques capable to jointly exploit the information of panchromatic and hyperspectral bands acquired satellites sensors; 4) classification of multi-temporal series of VHR or hyperspectral images with active learning and domain adaptation techniques for an automatic update of land-cover maps.