



ELSEVIER

Available online at www.sciencedirect.com ScienceDirect

Procedia Earth and Planetary Science 1 (2009) 1165–1172

Procedia Earth
and Planetary
Sciencewww.elsevier.com/locate/procediaThe 6th International Conference on Mining Science & Technology

Classification of quickbird image with maximal mutual information feature selection and support vector machine

Wu Bo^{a,*}, Xiong Zhu-guo^b, Chen Yun-zhi^a, Zhao Yin-di^c^a*Spatial Information Research Center of Fujian Province, Fuzhou University, Fuzhou 350002, China*^b*School of Earth Science and Survey Mapping, East China University of Technology, Fuzhou 344000, China*^c*School of Environmental Science and Spatial Informatics, China University of Mining and Technology, Xuzhou 221008, China*

Abstract

This paper presents a method to select optimal feature subset from object-orientated image segmentation according to the maximal mutual information to improve classification accuracy of high spatial resolution imagery over urban area. The proposed method is a three-step classification routine that involves the integration of 1) image segmentation with eCognition software, 2) feature selection by maximal mutual information criterion, and 3) support vector machine for classification. Experiment is conducted on Quick-Bird image in Fuzhou city. Furthermore, the proposed method with the well known feature selection methods, namely Tabu greedy search algorithm and fisher discriminate analysis, are evaluated and compared. The experiment shows that the mean error ratio significantly decreases with feature selection. It also demonstrates that the proposed maximal mutual information feature selection with support vector machine classifier significantly outperforms the classification method accompanied with eCognition platform in terms of Z test.

Keywords: maximal mutual information; feature selection; object-oriented classification; high spatial resolution image

1. Introduction

Obtaining detailed information about urban land use and environment is an important issue for local governments to seek better planning and management due to rapid development in urban area. High spatial resolution image has been proven to be a cost-effective method to monitor urban change information as it can economically provide rich detailed structural information for land cover/land use mapping in complex settings [1]. However, compared with middle spatial resolution image, such data provides detailed geometrical, structural and textural feature information, as well as creates higher spectral variance within each class corresponding to land-cover units owing to its fewer bands, which decreases their spectral separability and results in lower classification accuracy with traditional classified methods where only spectral information was used [2]. For instance, spectral characteristics of urban land cover classes such as road surfaces, parking lots, and open areas are so similar that they cannot be separated

* Corresponding author. Tel.: +86-591-87894818; fax: +86-591-87892536.

E-mail address: wavelet778@sohu.com.

correctly. As a result, there is an increased interest in incorporating geometrical or textural information with object-based methods in image classification [3-5].

The prominent advantage of object-oriented classification is that different shape and texture characteristics of objects can easily be calculated on the segments, as well as contextual information to improve the classification accuracy. Therefore, urban objects may be recognized as distinct blocks, and algorithms are developed based on “per-object” segmentation rather than “per-pixel” classification. Objects in urban area are possible to be separated by shape and texture in addition to spectral information. Many endeavors on object-oriented classification have been made and achieved great success. For instance, in [2] roofs and roads are discriminated using textural and shape information. Complex multi-scale frameworks have been developed in time for combining all the features and improving high spatial resolution image segmentation in urban areas [4]. A well known commercial software platform called eCognition [3] has also been developed specially for multi-scale object segmentation and classification.

While many methods for extracting different shape and textural characteristics of objects from remote sensing image on the segments have been developed in the past decades [6], few precise procedures for automatically selecting those features those are significant and thus useful for a particular classification problem are available [7]. The widely used manual methods for selecting significant features subset by expert experiments cannot ensure optimizing ones, and brings out a large of tedious work. Consequently, optimizing performance of the classification process entails the automatic selection of the most appropriate features. This paper presents a method to select good features from object-oriented image segmentation according to the maximal mutual information criterion and evaluated the capabilities of SVM classifier.

2. Methodology

The flowchart of the proposed method is presented in Fig.1. The distinctive merit from traditional object-based classification is that the proposed method closely integrated with eCognition software and injected automated feature selection procedure before object identification, where the subset features are generally manually selected. The proposed method includes three key steps for classification of high spatial resolution image that involves the integration of 1) multi-scale image segmentation, 2) automatic feature selection by some criterions, and 3) support vector machine for classification.

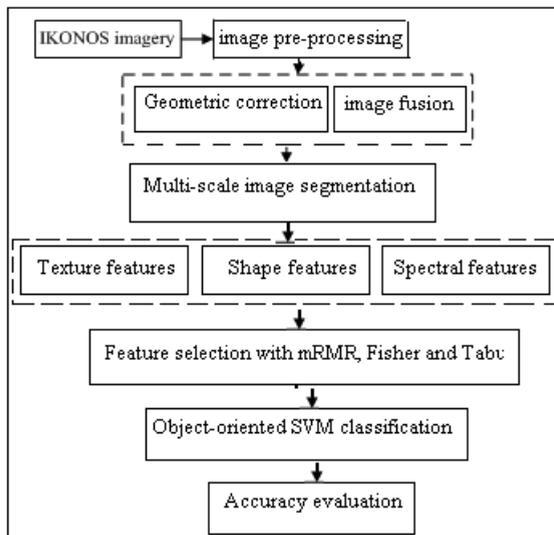


Fig. 1. Flowing chart for the proposed method

2.1. Object-oriented segmentation

We use commercial software eCognition 5.0 [3] for the purpose of object-oriented image segmentation. The

segmentation method is based on Fractal Network Evolution Approach (FNEA) algorithm, which is a bottom-up region merging technique by incorporating both spectral and spatial information. The main procedures are as followings: Random selecting a pixel, this pixel and its neighborhood in the minimum heterogeneity area are merged into a separate image object. Once mergence process is over, objects generated by the last mergence are used as basic units and calculates the heterogeneity of the object with its adjacent objects. Then, that adjacent smaller image objects in the defined heterogeneity are merged into bigger ones in a larger scale. The process will continue until there are no possible image objects to merge on the user specified scale. Moreover, local mutual optimal adaptive criterion is used to ensure that heterogeneity of every merging result is the minimum among all possible merging schemes. In application, the scale is a key parameter which controls the homogeneity size and shape. Therefore, in order to obtain reasonable segmentation results, appropriate scale parameter by different classification target is necessary.

2.2. Feature selection based on maximal entropy

Once objects are obtained after the image segmentation being conducted, we can then easily calculate different shape and texture characteristics of objects from the segmented image, such as spectral mean value, shape index, homogeneity and textural entropy, et al. With different scales and methods, a large number of image features will be obtained. Since not all these features are useful for classification, identifying the most characterizing features of the observed data, i.e. feature selection is critical to minimize the classification error. We adopt widely used maximal mutual information criterion for feature selection purpose in this paper, owing to its effectiveness and robustness.

Let two random x and y to be multidimensional variables, their mutual information is defined in terms of their probabilistic density functions $p(x)$, $p(y)$, and $p(x, y)$,

$$I(x, y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \quad (1)$$

The purpose of feature selection is to find a feature set S with m features, which jointly have the largest dependency on the target class c .

$$\max D(S, c) = I(\{x_i, i = 1, 2, \dots, m\}; c) \quad (2)$$

One difficult problem for this feature selection method is that jointly probabilistic density function $p(x,y)$ to be applied to seek the best subset among all the possible feature subsets, which results in a considerably high computational complexity. To circumvent the problem, an alternative suboptimal feature selection method based on maximal relevance and minimal redundancy criterion (mRMR) has been proposed by Peng et al. [8], which provides more practical solutions in terms of computational complexity although it cannot promise that the finally selected features subset is globally optimal. This method involves the integration of maximal relevance and minimal redundancy two aspects. Maximal relevance is to search features satisfying (3), which approximates $D(S,c)$ in (2) with the mean value of all mutual information values between individual feature x and class c .

$$\max V = \frac{1}{|S|^2} \sum_{i \in S} I(c, x_i) \quad (3)$$

In this way, all potential features having large mutual information values are possible to be selected even if they are highly depend on each other. Therefore, the selected feature subset is unavoidable highly redundancy, which decreases the discriminated capability. In order to overcome the shortcoming, the second criterion of minimal redundancy condition is applied to select mutually exclusive features.

$$\min W = \frac{1}{|S|^2} \sum_{i,j \in S} I(x_i, y_j) \quad (4)$$

The criterion combing the above two constraints are to define the operator to combine V and W and consider the following simplest form to optimal V and W simultaneously:

$$\max \phi = (V - W) \quad (5)$$

To unload computation of selecting features by (5), an incremental search approach is developed to find the near-optimal features defined by (6). Suppose the feature set with $m-1$ features are obtained, the task is to select the m th feature from the exclusive set. This is done by selecting the feature that maximizes (6). The respective incremental algorithm optimizes the following condition:

$$\max_{x_j \in X - S_{m-1}} [I(x_j, x_i) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j, x_i)] \quad (6)$$

These optimizations can be computed efficiently in $O(|S| m)$ complexity. As a result, we can obtain the ranked features rapidly even if the dimension of features is possible very high.

2.3. Support vector machine classifier

Support vector machine (SVM) is used for the purpose of evaluating the proposed feature selection method and classification. SVM is a relatively new and promising classification method. There are three prominent properties of SVM with respect to our research which can be summarized as follows: 1) high classification accuracies and very good generalization capabilities with respect to other traditional classifiers; 2) capability to address classification problem in which no explicit parametric models on the distribution of information classes are assumed; 3) possibility of defining nonlinear decision boundaries by implicitly mapping the available observations into high dimensional space. Because classification of high spatial resolution image involves high dimensional feature vectors that do not satisfied Gaussian normal distribution any longer. Therefore, SVM classifier is appropriate for the purpose of urban image classification.

Suppose we have N available training data points $\{(x_i, y_i)\}_{i=1}^N$, where $\{x_i\}_{i=1}^N$ is a subset of N features drawn from the remote sensing being classified and $\{y_i\}_{i=1}^N$, $y_i = \pm 1$ is the set associated with the true labels. We would like to learn a linear hyper-plane classifier, $f(x) = \text{sgn}(\mathbf{w} \cdot \mathbf{x} + b) = 0$ where \mathbf{x} represents a generic sample, \mathbf{w} is a vector normal to h , and b is a constant such that $b/\|\mathbf{w}\|$ represents the distance of h from the origin. The distance between the two hyper-planes $\mathbf{w} \cdot \mathbf{x} + b = \pm 1$ parallel to h is so called margin. Generally, the larger is the margin, the higher is expected to be the generalization capability of the classifier. Consequently, the objective of SVM is to minimize the following optimization problem:

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i \quad (7)$$

Subject to $y_i(\mathbf{w} \cdot \mathbf{x} + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$, for $i=1, 2, \dots, N$, where ξ_i 's are slack variables allowing for predicted error and C is the associated penalization parameter, which permits to tune the generalization capability. Since direct solving the above inequality constraints is difficult, the above optimal problem can be transfer to Wolfe dual problem as following:

$$\left\{ \begin{array}{l} \max \quad \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N y_i y_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j \\ \sum_{i=1}^N y_i \alpha_i = 0 \\ 0 \leq \alpha_i \leq C \end{array} \right. \quad (8)$$

where the coefficients α_i are referred as Lagrange multipliers. In most practical applications, the two classes cannot be linearly separated. To extend the linear learning machine to work with non-linear cases, SVMs use a kernel function to map the non-linearly separable classes from the input space to a higher dimensional feature space, in which the non-linearly separable classes can be separated by a linear optimal hyper-plane. As a consequence, the

inner product between two patterns \mathbf{x}_i and \mathbf{x}_j in (8) becomes $\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$. To avoid considering the mapping explicitly, it is possible to exploit kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$, which ensures that the objective function is convex. There are many kernel functions that can be used in SVMs. The most commonly used kernels include the polynomial kernel and radial basis function kernel. After obtaining the optimal values of the multipliers with some optimization techniques, for any given sample \mathbf{x} , the predicted label becomes

$$\hat{y} \equiv \text{sgn}[f(\mathbf{x})] = \text{sgn}\left[\sum_{i=1}^N y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b\right] \quad (9)$$

Note that standard SVM is for 2 classes. For multiclass, we can construct a multiclass classifier by combining binary classifier using one–against–one or one–against–all strategy.

3. Experiment results

3.1. Experimental data

To validate the effectiveness of the proposed methodologies, a subset image with 792×1133 pixels chosen from Quickbird of Fuzhou is used. The image was acquired in June 2003, collecting panchromatic images with 0.6-meter resolution and multispectral imagery with 2.4-resolution. According to the ground truth, the test site contains typical parcels in urban area with different types of roads, vegetation patches (e.g. road green belts, four-side trees, plantation), water, building areas, open surfaces and shadow regions. Panchromatic and multispectral images were fused in order to make full use of the spectral and spatial information, and finally the 0.6m multispectral image were obtained. And then the fused image was segmented with using eCognition software. The segmentation parameters are elaborately determined by trying. Figure 2 illustrates the segmented results taking the fused image as background.



Fig. 2. Display the segmented and fused image

3.2. Extracting object features

Various object features can be extracted from the segmented image. Generally, image spectral or gray information provides the most important measure features to identify ground object in remotely sensed imagery, spectral features, such as spectral mean value, standard deviations (SD), Brightness and max-min difference et al, are hence firstly extracted from segmented object. Image characteristics such as shape, pattern and texture are the possible most important features used in visual interpretation of aerial photographs and high spatial resolution imagery [7]. Measured textural parameters should therefore be considered as a potentially important feature source for urban mapping purposes. We abstract the widely used texture features based on co-occurrence matrices of the image gray level, and then Contrast, Homogeneity, Dissimilarity, and so on are calculated. In addition, Object shape is obtained by calculating object length/width proportion shape index and area et al. Moreover, to separate vegetation from other materials, normalized difference vegetable index (NDVI) is also extracted.

Table 1 summarizes the object features into four categories, including spectral, shape, texture and vegetable index features, which add up to 65 features.

Table 1. Summarized object features extracted from image

Feature category	Object features	Number of features
Spectral	Mean	5
	Band rate	5
	SD	5
	Brightness	1
	Max-Diff	1
	Contrast	5
	Homogeneity	5
Feature	SD	5
	Mean	5
	Dissimilarity	5
	Entropy	5
	ASM	5
	Area	1
	Length	1
	Width	1
	Length/Width	1
	Shape index	1
Shape	Density	1
	Compactness	1
	coast	1
	Aspect	1
	Board index	1
	ASM	1
	NDVI	1
VI	1	
Total		64

3.3. SVM Classification and evaluation

For classification of such a high dimensional data set, feature selection by maximal mutual information method aforementioned is used. This selected feature set for the dataset is tested using SVM classifier introduced in section 2. Gaussian Radial Basis Function (RBF) kernel, were used because they have been shown to be effective in ill-posed classification problems [9]. The parameters of SVM were selected according to a 3-fold cross validation (CV) procedure. The training data was randomly split into 3 mutually exclusive subsets of equal size and used to train a SVM classifier modeled with the predefined values. Each time, we kept one of the subsets from training and used it only to obtain an estimate of the classification accuracy. The 3 resultant estimates were averaged to yield the classification error rate for the SVM classifier. In order to demonstrate the proposed method, two other feature selection algorithms, i.e. greedy search algorithm and fisher discriminate analysis were also implemented for the same data.

Figure 3 shows their classification error rates with respect to selected number of features, respectively. It is clearly that mRMR approach achieves the lowest error rate by 14.9% with minimum 12 features, while the other two methods achieve their best recognized accuracies at 17 features. Both of their minimum error rates are above 16.2%. Figure 3 also shows that the classification error rate is 18.7% if all features are conducted, severely worse than those of feature selection method. Therefore, the experiment apparently demonstrates the importance of feature selection

in oriented-object classification.

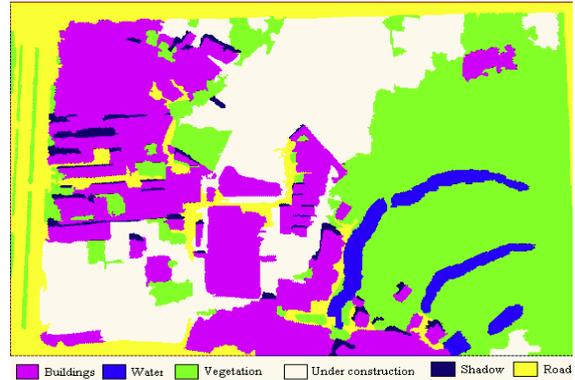
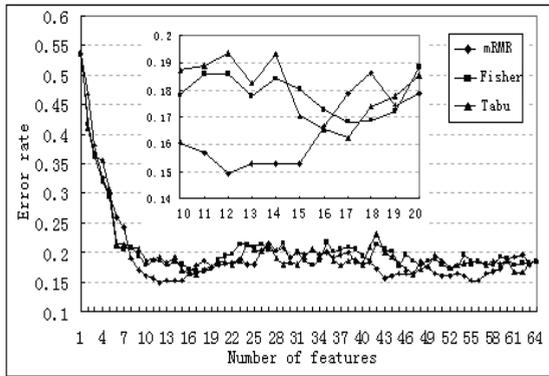


Fig. 3. classification error rates with respect to selected number of features

Fig. 4. Final classification map

Figure 4 presents the final classified map with SVM. Compared with ground truth, most of the results of figure 4 are in line with actual conditions, indicates the classification accuracy of this image is comparable high. The main classification errors result from the image segmentation, where some objects are inappropriate segmented. To further demonstrate the advantage of automatic feature selection, this image was also classified using manually chosen features and classified in eCognition software. In this situation, the final overall accuracy is 82.8%.

We utilized Z test[10] to assess whether the performance of mRMR is over traditional way significantly, This test is based on the standardized normal test statistic,

$$Z = \frac{c_{12} - c_{21}}{\sqrt{c_{12} + c_{21}}} \tag{10}$$

Where c_{12} denotes the number of samples classified correctly and wrongly by the 1st and 2nd models, respectively. Accordingly, c_{12} and c_{21} are the counts of classified samples on which the considered 1st and 2nd models disagree. A lower prediction error (higher accuracy) is identified by the sign on Z. A negative sign indicates that the results from c_{12} are more accurate than the results from model c_{12} . At the commonly used 5% level of significance, the difference in the accuracies between the 1st and 2nd models is evaluated to be statistically significant if $|Z| > 1.96$.

The Z value for the two methods based on overall accuracy is -3.31, indicating that the difference in the accuracies achieved by the two methods is statistically significant. In other words, automatic feature selection with mRMR can improve classification accuracy significantly.

4. Conclusions

- 1) Experimental results show that feature selection process in object-based classification can not only reduce the number of features greatly, but also improve the classification accuracy significantly.
- 2) The mRMR is the best feature selection method, compared with the other two widely used algorithms, namely Tabu greedy search and fisher discriminate analysis algorithm.

Acknowledgements

Financial support for this work provided by the Key Science Fund Project of Fujian province under the grant No. 2007Y0021; the High Educational Cultivated Fund Significant Science and Technology project of ministry of education under the grant No. 706037 are gratefully acknowledged; and Fujian Provincial Natural Science Fund under the grant No. 2009J01213

References

- [1] L. Bruzzone and M. Marconcini, Toward the Automatic Updating of Land Cover Maps by a Domain-adaptation SVM Classifier and a Circular Validation Strategy, *IEEE Transactions on Geoscience and Remote Sensing*. 47 (2009) 1108-1122.
- [2] L. Zhang, X. Huang, B. Huang and P. Li, A Ppixel Shape Index Coupled with Spectral Information for Classification of High Spatial Resolution Remotely Sensed Imagery, *IEEE Transactions on Geoscience and Remote Sensing*. 44 (2006) 2950–2961.
- [3] Definiens Imaging. *eCognition Professional User Guide 5*. Munich, Germany, 2003
- [4] S. Myint, N. Lam and J.M. Tylor, Wavelets for Urban Spatial Feature Discrimination: Comparisons with Fractal, Spatial Autocorrelation, and Spatial Co-occurrence Approaches, *Photogrammetric Engineering and Remote Sensing*. 70 (2004) 803-812.
- [5] L. Bruzzone and L. Carlin, A Multilevel Context-based System for Classification of Very High Spatial Resolution Images, *IEEE Transactions on Geoscience and Remote Sensing*. 44 (2006) 2587-2600.
- [6] G.J. Hay, T. Blaschke and D.J. Marceau, A Comparison of Three Image-object Methods for the Multiscale Analysis of Landscape Structure, *ISPRS Journal of photogrammetry & remote sensing*. 57 (2003) 327-345.
- [7] M.B.Fricke, P.C. Lieven and R. Robet, Feature Selection by Genetic Algorithms in Object-based Classification of IKONS Imagery for Forest Mapping in Flanders Belgium, *Remote Sensing of Environment*. 110 (2007) 476-487.
- [8] H. Peng, F. Long and C.F. Ding, Feature Selection Based on Mutual Information: Criteria of Max-dependency, Max-relevance, and Min-redundancy, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 27 (2005) 1226-1238.
- [9] O. Chapelle, V.Vapnik, O. Bousquet and S. Mukerjee, Choosing Multiple Parameters for sSupport Vector Machines, *Machine Learning*. 46 (2002) 131-159.
- [10] G.M. Foody, Thematic Map Comparison: Evaluating the Statistical Significance of Differences in Classification Accurac, *Photogrammetric Engineering and Remote Sensing*. 70 (2004) 627-633.