International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 5 1214 – 1219

Satellite Data Classification Based On Support Vector Machine, Rough Sets Theory & Rough-SVM

Lekha Bhambu
(Phd Scholar)
Guru Kashi University, Talwandi Saheb
lbhambhu@gmail.com

Dr. Dinesh Kumar
(Associate Professor)
Guru Kashi University, Talwandi Saheb
kdinesh.gku@gmail.com

Abstract: As Classification is becoming one of the most crucial tasks for various applications. Text categorization, tone recognition, image classification, micro-array gene expression are the few examples of such kind of applications. The supervised classification techniques are mostly based on traditional statistics capable of providing good results when sample size seems to tend to infinity. But in practice, only finite samples can be acquired. In this paper, an innovative learning technique, Rough Support Vector Machine (SVM), is employed on Satellite Data multi class. SVM Initiated in the early 90's, a powerful machine technique amplified from arithmetical learning led to an outburst of interest in machine learning and have made noteworthy achievement in some field as SVM technique does not agonize the boundaries of data dimensionality and limited samples [1] & [2].

In our investigation, as the support vectors, classification are gathered by learning from the training samples are very perilous. In this paper, using various kernel functions for satellite data samples relative outcomes explained.

Keywords: *Classification, SVM, RSES, Kernel functions, Grid search, Rule Base Classifier.*

*****

I. INTRODUCTION

Firstly Vapnik offers the Support Vector Machine and machine learning research community inirting the shrill concern. [2]. There are many several parts in other countries that reported the SVM (support vector machines) generally are capable of excellent accuracy performance. Generally used for accuracy than the other data classification algorithms. However a wide range of real world problems such as text categorization, hand-written digit recognition, tone recognition, image classification and object detection, micro-array gene expression data analysis, data classification. Generally it shown that Sims is consistently superior to other supervised methods. Some datasets, in performances of SVM are very sensitive to the usage of cost parameter and kernel parameters are for set. Since the user higherly needs to conduct extended cross validation in order to figure out the optimal parameter setting. Model selection recommended by selected conduct process. This model selection is very time consuming. There are many generated people who experimented with a number of parameters associated with the use of the SVM algorithm that can impressive results. The choice of kernel functions, or standard deviation of the Gaussian kernel, mainly relative weights associated with slack variables to account for the non-uniform distribution of labeled data, and the number of these training examples are included in these parameters.

For example, we recommended satellite data which have different features, classes, number of training data and different number of tested data. These all data taken from RSES data set and

http://www.ics.uci.edu/~mlearn/MLRepository.html[5]

A new mathematic tool is rough set deals with un-integrality and indefinite knowledge. It can effectively used for analyzed deal with all kinds of fuzzy, conflicting and incomplete information, and finds out the connotative knowledge from it, and view out its underlying rules. It was first generated by Z.Pawlak, a Polish mathematician, in 1982. In recent several years, rough set theory is widely process for the application in the fields of data mining and artificial intelligence.

This paper is organized as follows. In next section, we introduce some related background including some basic concepts of SVM, kernel function selection, and model selection (parameters selection) of SVM. In Section 3 we introduce about Rough Set Method. In Section 4, we detail all experiments results. Finally, we have some conclusions in Section 5 and References in section 6.

II. SUPPORT VECTOR MACHINE

In this section some basic concepts of SVM, different kernel function, and model selection (parameters selection) of SVM are introduced [2] [3][17] .

Overview of SVM

SVM is a set of related supervised learning methods and we used this method for classification and regression [2].

They belong to a group of linear classification. As SVM is considered as an important cast simultaneously to reduce the empirical classification error and enhance the geometric margin. This SVM called Maximum Margin Classifiers. SVM is based on the Structural risk Minimization (SRM). SVM mapped input vector to a higher dimensional space where as a highest overview of a separating hyperplane is constructed. Two parallel hyperplanes are constructed between each side of the hyperplane that separates data. The separating hyperplane is maximize the distance between the two parallel hyperplanes. An assumption is that the larger of the margin or distance between these parallel hyperplanes in the better results of the generalization error of the classifiers.[2].

We consider the data points in the form of

$\{(x_1,y_1),(x_2,y_2),(x_3,y_3),(x_4,y_4)\ldots\ldots\ldots,(x_n, y_n)\}$.

Where y_n=1 / -1 , a constant denoting the class to which that point x_n belongs. n = number of sample. Each x_n is p-dimensional real vector. As the scaling is important to guard against variables with larger of the variance. We analyze this Training data, by means of the separating hyperplane , which takes in the form of

$$w \cdot x + b = o \qquad \text{--------(1)}$$

Where b is scalar and w is p-dimensional Vector.

The vector point of w is perpendicular and it separates the hyperplane. B is the offset parameter and it allows us to increase the margin. Absence of b, the hyperplane is forced to pass through the origin, and restricting the solution. As we are interesting in the maximum margin, we takes SVM and the parallel hyperplanes. As Parallel hyperplanes can be described by equation i.e.

$$w.x + b = 1$$

$$w.x + b = -1$$

If the training datasets are linearly separable, as we can select hyperplanes so that there are no points between them and then try to increase their distance. By geometry, We can find the distance between the hyperplane is $2 / |w|$. So we want to minimize $|w|$. To excite data points, we need to ensure that for all I either

$$w \cdot x_i - b \geq 1 \quad \text{or} \quad w. x_i - b \leq -1$$

Or this can be written as

$$y_i (w. x_i - b) \geq 1 \quad , \ 1 \leq i \leq n \qquad \text{----------(2)}$$

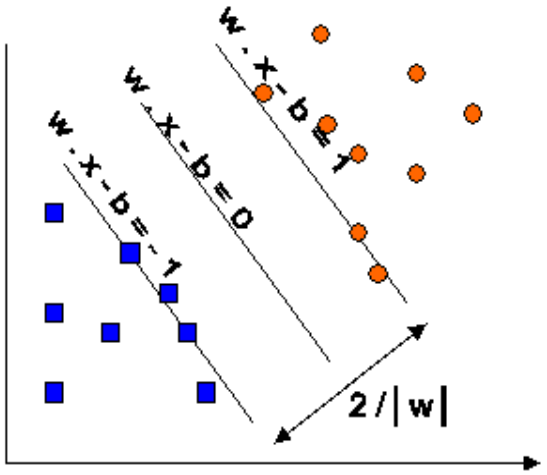

Figure.1 Maximum margin hyperplanes for a SVM trained with samples from two class

Samples along the hyperplanes are called Support Vectors (SVs). As separating hyperplane with the largest margin defined by $M = 2 / |w|$ that is specifies support vectors training data points closest to it. which satisfies

$$y_j [w^T . x_j + b] = 1 \quad , i = 1 \qquad \text{----------(3)}$$

Optimal canonical hyperplane (OCH) is a canonical hyperplane having a maximum margin. For all the data , OCH should satisfy the following constraints

$$y_i[w^T . x_i + b] \geq 1 \quad ; \ i = 1,2\ldots\ldots l \qquad \text{----------(4)}$$

Where l is Number of Training data point. In order to find the optimal separating hyperplane having a maximum margin, A learning machine should be minimize $\|w\|^2$ subject to the inequality constraints as

$$y_i [w^T . x_i + b] \geq 1 \quad ; \ i = 1,2\ldots\ldots l$$

This optimization problem solved by the saddle points of the Lagrange's Function

$$L_P = L_{(w, b, \alpha)} = 1/2 \|w\|2 - \sum_{i=1} \alpha_i \ (y_i (w^T x_i + b) - 1)$$

$$L = 1/2 \ w^T w - \sum_{i=1} \alpha_i \ (_{yi}(w^T x_i + b) - 1) \qquad \text{---------(5)}$$

Where αi a Lagrange's multiplier .The search for an optimal saddle points (w_0, b_0, α_0) is necessary because Lagrange's must be minimized w.r.t w and b and has to be maximized w.r.t. nonnegative αi ($\alpha_i \geq 0$). This problem can be solved either in primal form (which is in the form of w & b) or in a dual form (which is in the form of α_i).Equation number (4) and (5) are convex and KKT conditions, which are necessary and sufficient conditions for a maximum of

equation (4). Partially differentiate equation (5) w.r.t. saddle points (w_0, b_0, α_0).

. $\partial L / \partial w_0 = 0$

i .e $w_0 = \sum_{i=1} \alpha_i \ y_i \ x_i$ ----------(6)

and $\partial L / \partial b_0 = 0$

i .e $\sum_{i=1} \alpha_i \ y_i = 0$ ----------(7)

Substituting equation (6) and (7) in equation (5) . we change the primal form into dual form.

$$L_d(\alpha) = \sum \alpha_i - 1/2 \sum \alpha_i \alpha_j \ y_i y_j \ x_i^T x_j \text{ ----------(8)}$$

AS optimal hyperplane should be selected so a dual lagrangian (L_d) has to be maximized w.r.t. nonnegative αi (i.e. α_i must be in the nonnegative quadrant) and w.r.t. the equality constraints as follows:

$$\alpha_i \geq 0 \quad , \ i = 1,2\ldots.l$$

$$\sum_{i=1} \alpha_i \ y_i = 0$$

Note that the dual Lagrangian $L_d(\alpha)$ is expressed in terms of training data and depends only on the scalar products of input patterns ($x_i^T x_j$). More detailed information on SVM can be found in Reference no.[1]&[2]

Kernel Selection of SVM

The function Φ is used for mapping the Training vectors x_i into a higher (may be infinite) dimensional space . Then SVM finds a linear separating hyperplane with the maximal margin in this higher dimension space .C > 0 is the penalty parameter of the error term. Furthermore, $K(x_i , x_j) \equiv \Phi(x_i)^T \Phi(x_j)$ is called the kernel function[2]. There are many kernel functions in SVM, so how to select a good kernel function is also a research issue. However, for general purposes, there are some popular kernel functions[2]&[3]:

- linear kernel:
  $K(x_i , x_j) = x_i^T x_j$.
- Polynomial kernel:
  $K(x_i , x_j) = (\gamma \ x_i^T x_j + r)^d \quad , \quad \gamma > 0$
- RBF kernel:
  $K(x_i , x_j) = \exp(-\gamma \ \| x_i - x_j \|^2) \ , \quad \gamma > 0$
- Sigmoid kernel:
  $K(x_i , x_j) = \tan h(\gamma \ x_i^T x_j + r)$.

Here, γ, r and d are kernel parameters. In these popular kernel functions, RBF is the main kernel function because of following reasons [2]:

- The RBF kernel nonlinearly maps samples into a higher dimensional space unlike to linear kernel.
- The RBF kernel has less hyperparameters than the polynomial kernel.

The RBF kernel has less numerical difficulties.

Model Selection of SVM

Model selection is very important issue in SVM. Mostly, SVM have shown good performance in data classification in the mainly part of success depends on the tuning point of several parameters which affects the generalization error. We often call this parameter tuning procedure as the model selection. If you use the linear SVM, you only need to tune the cost parameter C. Basically**,** linear SVM are often applied to linearly separable problems. Many problems are non-linearly separable. For example, Satellite data and Shuttle data are not linearly separable. Therefore, we often apply nonlinear kernel to solve classification problems, so we need to select the cost parameter (C) and kernel parameters(γ , d)[4]&[5].

We usually use the grid-search method in cross validation to select the best parameter set. Then apply this parameter set to the training data set and then we get the classifier. After that, For classifying the testing of dataset to get the generalization accuracy uses the classifier.

1. INTRODUCTION OF ROUGH SET

Rough set is a new mathematic tool which is capable to deals with Un-integrality and indefinite knowledge. It not only effectively analyzes and deal with types of fuzzy but also the conflicting and Incomplete information and still finds out the connotative Knowledge from it and reveals its underlying rules. It was first made by Z.Pawlak, a Polish mathematician, in 1982. In few years, rough set theory is widely emphasized for the application in the fields of data mining and artificial intelligence[7] [8] [13].

The basic definitions of rough set

Let S be an information system formed of 4 elements

S = (U, Q, V, f) where

U - is a finite set of objects

Q - is a finite set of attributes

V- is a finite set of values of the attributes

f- is the information function so that:

f : U × Q - V.

Let P be a subset of *Q, P* $\subseteq$ *Q,* i.e. a subset of attributes. The indiscernibility relation noted by IND(P) is a relation defined as follows

IND(P) = {< x, y > ∈ U × U: f(x, a) = f(y, a), for all a ∈ P}

If < *x, y* > ∈ *IND(P),* then we can say that x and y are indiscernible for the subset of P attributes. *U/IND(P)* indicate the object sets that are indiscernible for the subset of P attributes.

U / IND(P) = { U_1, U_2, …….U_m }

Where U_i ∈ *U, i = 1 to m* is a set of indiscernible objects for the subset of P attributes and *Ui* ∩ *Uj* = *Φ,*

i ,j = 1to m and i ≠ j. *Ui* can be also called the equivalency class for the indiscernibility relation. For *X* $\subseteq$ *U* and P inferior approximation P_1 and superior approximation P^1 are defined as follows

$P_1(X) = U\{Y \in U/ IND(P): Y \subseteq Xl\}$

$P^1(X = U\{Y \in U / INE(P): Y \cap X \neq \Phi\}$

In feature selection Rough Set Theory is successfully used. It is based on finding a reduct from the original set of attributes. Now Data mining algorithms will not process the original set of attributes, but this reduct that will be equivalent with the original set. The set of attributes Q from the informational system *S = (U, Q, V, f)* can be divided into two subsets: C and D, so that *C* $\subset$ *Q, D* $\subset$ *Q, C* ∩ *D* = *Φ* . Subset C will contain the attributes of condition, while subset D those of decision. Equivalency classes *U/IND(C) is condition classes* and *U/IND(D) is* decision classes.

The degree of dependency is the set of attributes of decision D as compared to the set of attributes of condition C is marked with γc (D) and is defined by

$$\gamma_C(D) = \frac{|POS_C(D)|}{|U|}, \quad 0 \le \gamma_C(D) \le 1$$

$$POS_C(D) = \bigcup_{X \in U / IND(D)} \underline{C}X$$

POSC (D) contains the objects from U which can be classified as belonging to one of the classes of equivalency U/IND(D), using only the attributes in C. if $\gamma_c(D) = 1$ then C determines D functionally. Data set U is called consistent if $\gamma_c(D) = 1$. $POS_C(D)$ is called the positive region of decision classes U/IND(D), bearing in mind the attributes of condition from C.

Subset *R* $\subset$ *C* is a D-reduct of C if POS_R (D) = POS_C(D) and R has no R' subset, R' $\subset$ R so that $POS_{R'}$.(D) = POS_R(D) . Namely, a reduct is a minimal set of attributes that maintains the positive region of decision classes U/IND(D) bearing in mind the attributes of condition from C. Each reduct has the property that no attribute can be extracted from it without modifying the relation of indiscernibility. For the set of attributes C there might exist several reducts.

The set of attributes that belongs to the intersection of all reducts of C set is called the core of C.

$$CORE(C) = \bigcap_{Re REDUCT(C)} R$$

An attribute is a indispensable for C if POS_C (D) ≠ $POS_{C[a]}$ (D). The core of C is the union of all indispensable attributes in C. The core has two equivalent definitions. More detailed information on RSES can be found in .[1]&[2].

2. ROUGH-SVM

Rough Sets Theory is an efficient tool in processing imprecise or vague concepts. There are following advantages

- It is based only on the original data and does not need any external information, unlike probability in statistics or grade of membership in the Fuzzy set theory.
- It can reduce conditional attribute and eliminate redundant information, but not reduce any effective information.
- It is a tool suitable for analyzing not only quantitative attributes but also qualitative ones.

A type of support vector machine classification system based on the Rough Sets pre-process is presented in this section. Given a training sample set, we firstly discrete them if the sample attribute values are continuous, and we can get a minimal feature subset that fully describe the all concepts by attribute reduction, constructing a support vector classifier and finding a decision function *f (x) = (w . x) + b* . When given a test sample sets, we reduce the corresponding attributes and put into SVM classification system, then we can acquire the testing result. Figure II shows a flow diagram of ROUGH-SVM approach.

3. RESULTS OF EXPERIMENTS

The classification experiments are conducted on Satellite data.Satellite Data Set are Multi-spectral values of pixels in 3x3 neighborhoods in a satellite image, and the classification also associated with the central pixel in each neighborhood are included in databases. The mainly focus is to predict this classification, given the multi-spectral values. The class of a pixel is coded as a number in the sample database.

Many sources of information data available for a scene is the Landsat satellite data. The interpretation of a scene by integrating spatial data of diverse types. Resolutions including radar data multi-spectram and maps indicating topography as well as land use etc is expected to assume significant importance of data with the onset of an era characterized by integrative approaches for remote sensing as NASA's Earth Observing System is commencing this decade. Diverse data types are handled by well-equipped Existing statistical methods. This is not true for Landsat MSS data considered in isolation (as in this sample database). This data satisfies the significance requirements of being numerical and at a single resolution, and standard maximum-likelihood classification performs very well equipped. Frequently, for this data, Statistical approach is very interesting to compare by the performance of other methods.

Single frame of The Landsat MSS imagery has four digital images in the different spectral bands of the same scene. Two of these are in the visible region corresponding approximately to green and red regions of the visible spectrum and two are in the near infra-red region. Each pixel contains 8-bit binary word, with 0 corresponding to black and 255 to white. The higher resolution of a pixel is about 80m x 80m. Each image has 2340 x 3380 such kind of pixels.

The database is a small sub-area of a scene, having 82 x 100 pixels. Here each line of data corresponds to each 3x3 square neighborhood of pixels completely contained within 82x100 sub-area. Each line contains the pixel values in the four spectral bands (converted to ASCII) of each of the 9 pixels in the 3x3 neighborhood and a number indicating the classification label of the central pixel. The number is a code for the different classes i.e. “1” is red soil,” 2” is Cotton crop,”3” is grey soil,”4” is damp grey soil,”5” is soil with vegetation stubble,”6” is mixture class(all types present),”7” is very damp grey soil.

In this dataset there are no examples for class 6. The data isgiven in a random order. In each line of data the four different spectral values for the top-left pixel are given first followed by the four differentspectral values for the top-middle pixel and then those for the top-right pixel, and so on with the pixels read out in sequence left-to-right and top-to-bottom. Thus, the four spectral values for the central pixel are given by attributes i.e. 17, 18, 19 and 20. If you like you can use only these four attributes, while ignoring the others. This avoids the problem which arises when a 3x3 neighborhood straddles a boundary.

This data taken from http://www.ics.uci.edu/~mlearn/MLRepository.html and RSES data sets . In these experiments, we have done both methods on different data set. Firstly, Use LIBSVM with different kernel linear , polonomial , sigmoid and RBF[5]. RBF kernel is employed. Accordingly, there are two parameters, the RBF kernel parameter γ and the cost parameter C, to be set. Table 1 lists the main characteristics of the three datasets used in the experiments. This data sets is taken from the machine learning repository collection. In these experiments, 5-fold cross validation is conducted to determine the best value of different parameter C and γ .The combinations of (C, γ) is the most appropriate for the given data classification problem with respect to prediction accuracy. The value of (C , γ) for all data set are shown in Table 1. Second, RSES Tool set is used for data classification with all data set using different classifier technique as Rule Based classifier, Rule Based classifier with Discretization, K-NN classifier and LTF (Local Transfer Function) Classifier. For data classification some hardware used with all data set using different classifier technique as Rule Based classifier with Discretization, Rule Based classifier, K-NN classifier and Local Transfer Function Classifier. The hardware platform used in the experiments is a workstation with Pentium-IV-1GHz CPU, 256MB RAM, and the Windows XP(using MS-DOS Prompt).

The following three tables represents the different experiments results. Table 1 shows the best value of different RBF parameter value (C , γ) and cross validation rate with 5-fold cross validation using grid search method[5]&[6]. . Table 2 shows the Total execution time for satellite data to predict the accuracy in seconds. Table 3 shows percentage classification of satellite data using different classifier of rough set exploration system(RSES) & support vector machine using RBF kernel.

4. CONCLUSION

In this document, it is illustrated that there are relative outcomes through various kernel functions and the relative outcomes are reassured. In particular, as the selection of kernel function and best value of parameters for certain kernel and model selection are observed as very important for a specified quantity of data. Rough Support Vector Machine is classifying the satellite data with highest accuracy.

REFERENCES

[1] Boser, B. E., I. Guyon, and V. Vapnik (1992). A training algorithm for optimal margin classifiers . In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pages. 144 -152. ACM Press 1992.

[2] V. Vapnik. The Nature of Statistical Learning Theory . NY: Springer-Verlag. 1995.

[3] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A Practical Guide to Support Vector Classification . Department of Computer Science National Taiwan http://www.csie.ntu.edu.tw/~cjlin 2007

[4] [4] C.-W. Hsu and C. J. Lin. A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks, 13(2):415-425, 2002.

[5] Chang, C.-C. and C. J. Lin (2001). LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm .

[6] Li Maokuan, Cheng Yusheng, Zhao Honghai "Unlabeled data classification via SVM and k-means Clustering". Proceeding of the International Conference on Computer Graphics, Image and Visualization(CGIV04),2004 IEEE.

[7] Z. Pawlak, Rough sets and intelligent data analysis, Information Sciences 147 (2002) 1–12.

[8] RSES 2.2 User's Guide Warsaw University

[9] http://logic.mimuw.edu.pl/»rses ,January 19, 2005

[10] Eva Kovacs, Losif Ignat, "Reduct Equivalent Rule Induction Based On Rough Set Theory", Technical University of Cluj-Napoca.

[11] RSES Homepage http://logic.mimuw.edu.pl/»rses

[12] J. Komorowski; Z. Pawlak; L. Polkowski; A. Skowron: "Rough sets: A tutorial". In: S. K. Pal; A. Skowron (Eds.), Rough Fuzzy Hybridization: New Trend

[13] in Decision-Making Singapore: Springer-Verlag, 1999, 3-98, (1999).

[14] "Automatic Visual Inspection and Classification Based On Rough Sets and Neural Network", Meng-Xin Li, Cheng-Dong Wu, Yong Yue, 2003 IEEE.

[15] "Rough Set Based Classification rules generation for SARS Patients", Feng Honghai, Chen Guoshun, Wang Yufeng, Yang Bingru, Chen Yumei, 2005 IEEE.

[16] "An Email Classification Model Based on Rough Set Theory", Wenqing Zhao and Zili Zhang, 2005 IEEE.

[17] "Reduct Equivalent Rule Induction Based on Rough Set Theory", Eva Kovacs and losif Ignat, 2007 IEEE.

[18] "Recognition Method of Radar Signal Based on Rough Set and Support Vector Machine", Chen Ting Luo Jingqing1-4244-1372-9/07 c2007 IEEE.

[19] "Data classification using Support Vector Machine", Durgesh Srivastava & Lekha Bhambhu, Journal of Theoretical and Applied Information Technology.

Applications	Training Data	Testing Data	Best c and γ with five fold		Cross validation rate
			C	γ	
Satellite Data	4435	2000	2^1=2	2^1=2	91.725

Table 1: Best value using Grid Search method

Applications	Total Execution Time to Predict in second		
	SVM	RSES	Rough- SVM
Satellite data	747.49	85	1112.52

Table 2 : Execution Time in Seconds using SVM, Rough Set & ROUGH-SVM.

Applications	Training data	Testing data	Feature	No. Of Classes	Using SVM (with RBF kernel)	Using RSES with Different classifier				Rough SVM
						Rule Based Classifier	Rule Based Classifier with Discretization	K-NN Classifier	LTF Classifier	
Satellite Data	4435	2000	36	7	91.8	87.5	89.43	90.4	89.7	92.8

Table 3: Compare Rough-SVM with Rough Set Classifier & others