Contemporary Engineering Sciences, Vol. 3, 2010, no. 2, 77 - 86

Data Classification: A Rough - SVM Approach

D. K. Srivastava

Sr. Lecture, Deptt. of CSE/IT, BRCM CET, Bahal, Bhiwani, Haryana, India
dkumar.bit@gmail.com

K. S. Patnaik

Sr.Lecturer, Deptt of CSE, BIT, Mesra, Ranchi, 835215, India
ktosri@rediffmail.com

L. Bhambhu

Sr. Lecture, Deptt. of CSE/IT, BRCM CET, Bahal, Bhiwani, Haryana, India
lbhambhu@gmail.com

Abstract

Classification is one of the most important tasks for different applications. Most of the existing supervised classification methods are based on traditional statistics, which can provide ideal results when sample size is tending to infinity. However, only finite samples can be acquired in practice. SVM, a powerful machine method developed from statistical learning and has made significant achievement in some field. Introduced in the early 90's, they led to an explosion of interest in machine learning. The foundations of SVM have been developed by Vapnik and are gaining popularity in field of machine learning due to many attractive features and promising empirical performance. SVM method does not suffer the limitations of data dimensionality and limited samples.

This paper reports the introduction of Rough -SVM Approach based on the hybridization of SVM and Rough Set Exploration System (RSES). RSES is used to find reducts which then applied to SVM to obtain better classification results.

Keywords: Support Vector Machine (SVM), Rough Set Exploration System (RSES), Attribute Reduction, Kernel Functions, Classifications.

1. Introduction

The Support Vector Machine (SVM) was first proposed by Vapnik [2] and has since attracted a high degree of interest in the machine learning research community. Several recent studies have reported that the SVM generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. SVMs have been employed in a wide range of real world problems such as text categorization, tone recognition, micro-array gene expression and data classification. Rough Set Theory, which is developed by Z. Pawlak in 1982, is a new effective tool in dealing with vagueness and uncertainty information. Attribute reduction is one of the most important concepts. Irrelevant and redundant attributes are removed from the decision without any classification information loss [8]. This paper reports the use of RSES version 2.2 for attribute reduction and then application of SVM on new reduct data set for best classification results.

This paper is organized as follows. Section 2 describes some basic concepts of SVM. Section 3, highlights the concepts of Rough Set Theory section 4, gives the Introduction of A ROUGH-SVM Approach using SVM and RSES. Experimental results are described in Section 5. Finally, we have conclusion in Section 6.

2. Support Vector Machine

In this section we introduce some basic concepts of SVM, different kernel function, and model selection (parameters selection) of SVM.

A. Overview of SVM

First, we briefly describe some concepts of SVM. Given training data xi, i =1,...,n. in two classes, and a label vector y; such that yi є {1,-1}, the standard SVM formulation [6] is as follows:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i$$

Subject to

$$y_i \left(w^T \phi(x_i) + b \right) \geq 1 - \xi_i, \text{-------(1)}$$

$$\xi_i \geq 0, i = 1, \ldots, n.$$

If ф(xi)= xi , we often call (1) as the linear kernel SVM, which is mainly to solve the linearly separable problem. Unfortunately, many applications in the real world are not linearly separable problems. Accordingly, we use ф to map xi into a higher dimensional space, and then call (1) a non-linear SVM.

For a non-linear SVM, after mapping by ϕ, the number of variables of w can be very large or even infinite, so that it is very difficult to solve this problem from (1). As a result, people often solve the problem from the following dual formulation:

$$\min_{\alpha} \frac{1}{2} Q^T Q\alpha - e^T \alpha$$

Subject to $$y^T \alpha = 0, \qquad \text{---------------------} (2)$$

$$0 \le \alpha_i \le C, i = 1, ..., n.$$

where Q is an n × n positive semi-definite matrix with $Q_{ij} = y_i\, y_j\, \phi(x_i)^T \phi(x_j)$, and e is the vector with all 1 elements. Usually we call $K(x_i, x_j) = \phi\ (x_i)^T\ \phi(x_j)$ the kernel function. Some popular kernel functions are, for example, $e^{-\gamma}\|x_i - x_j\|^2$ (RBF), $(x_i^T, x_j/\ \gamma + \delta)^d$ (polynomial), $\tanh(\alpha\ x^T y + b)$ (hyperbolic tangent) etc., where γ, d and δ are kernel parameters. In addition, (2) is easier to be solved than (1) because the number of variables in (2) is the size of the training dataset, n, not the dimensionality of ϕ (x).
It can be shown that if a is an optimal solution of (2), then

$$w = \sum_{i=1}^{n} \alpha_i y_i \phi(x_i)$$

is the optimal solution of the primal (1). Then a decision function is written as

$$\operatorname{sgn}\left(w^T \phi(x) + b\right) = \operatorname{sgn}\left(\sum \alpha_i y_i K(x_i, x) + b\right)$$

That is, for a test vector x, if

$$\sum_{i-1}^{n} y_i \alpha_i \left(\phi(x_i)^T\ \phi(x)\right) + b \rangle 0$$

we classify it to be in the class 1. Otherwise, we think it is in the second class. Moreover, after (2) is solved with a solution α, the vectors for which $\alpha_i > 0$ are called support vectors. We can see that only support vectors will affect results in the prediction stage.

B. Kernel Selection of SVM

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: linear kernel, RBF kernel, polynomial kernel and hyperbolic tangent kernel. In these popular kernel functions, we often choose the RBF kernel function because of following reasons:

1. Some problems are not linearly separable, so we don't choose the linear kernel function.
2. We don't choose polynomial kernel function due to some numerical difficulties such as $(< 1)^d \rightarrow 0$, and $(>1)^d \rightarrow \infty$.
3. Hyperbolic tangent kernel is not well studied now, but it seems to behave like RBF kernel for certain parameters.

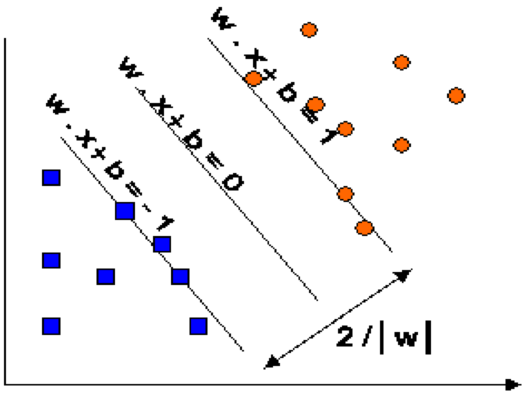


Fig.I Maximum margin hyper planes for a SVM trained with samples from two classes

C. Model Selection of SVM

Model selection is also an important issue in SVM. Recently, SVM have shown good performance in data classification. Its success depends on the tuning of several parameters which affect 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. Unfortunately, linear SVM are often applied to linearly separable problems. Many problems are non-linearly separable. For example, the problem shown in Figure 1 is obviously not linearly separable. Therefore, we often apply nonlinear kernel to solve classification problems, so we need to select the cost parameter and kernel parameters. As we discussed in the previous subsection, we often choose the RBF kernel function in general applications. In the RBF kernel function $K(x_i, x_j) = e^{-\gamma||x_i-x_j||^2}$, we need to select the cost parameter C and kernel parameter γ.

We usually use the grid-search method in cross valida-tion to select the best parameter set. That is, to do the cross validation in training dataset by trying different parameter combinations (often 15×15 = 225 combinations) to get the best one. Then apply this parameter set to the training dataset and then get the classifier. After that, use the classifier to classify the testing dataset to get the generalization accuracy.

3. Rough Set Theory

Z.Pawlak put the rough set theory mentioned in this paper forward in 1982. It is a new mathematic tool to deal with noise information, uncertain information, fuzzy information and incomplete information etc. It even needs no other a priori information except the data set concerning the problem under question. Meanwhile, the knowledge mined by rough set theory can be expressed and saved as rules, and we can apply these rules to reason out the result. Such a procedure is reliable and clear. Due to these advantages, rough set theory has been widely used in many fields such as artificial intelligent, data mining, decision-making and so on [8].

A. 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:

$f: U \times Q - V.$

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

$$IND (P) = \{< x, y > \in U \times U: f(x, a) = f(y, a), \text{for all } a \in P\}$$

If $< x, y > \in$ 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.

$$U / IND (P) = \{ U_1, U_2, \ldots\ldots U_m \}$$

Where $U_i \in U$, i = 1 to m is a set of indiscernible objects for the subset of P attributes and $U_i \cap Uj = \Phi$, i, j = 1to m and $i \neq j$. U_i 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

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

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

Rough Set Theory is successfully used in feature selection and is based on finding a reduct from the original set of attributes. Data mining algorithms will not run on the original set of attributes, but on this redact 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 $C \subset Q$, $D \subset Q$, $C \cap D = \Phi$. Subset C will contain the attributes of condition, while subset D those of decision. Equivalency classes U/IND(C) and U/IND (D) are called condition classes and decision classes

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

$$\gamma_c(D) = \frac{|POS_c(D|}{|U|} \quad , \quad 0 : \gamma_c(D) : 1$$

$$|POS_c\ (D)| = \underset{X \in U/IND\ (D)}{U} \subseteq X$$

POS_C (D) contains the objects from U that can be classified as belonging to one of the classes of equivalency U/IND (D), using only the attributes in C. If γ_c (D) = 1 then C determines D functionally. Data set U is called consistent if γ_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. More detailed information on RSES can be found in. [7] & [8].

B. Discretization

Data discretization is a widely used data transformation procedure, which involves finding cuts in the data set, which divide in the data into intervals. If the data contains noncategorical attributes, they must be discretized in order to produce effective rules. Values lying within an interval are then mapped to the same value. Doing this reduces the size of the attributes value set and ensures that the rules that are mined are not too specific. The cut points are applied to the training set with the same cuts applied to the test set.

C. Reducts

Now, Rough set theory is used to simplify the discretized decision table in order to obtain simpler classifications. One way to reduce the dimension of the data is to remove attributes whose removal will not destroy the indispensability relation. Attributes, which can be removed

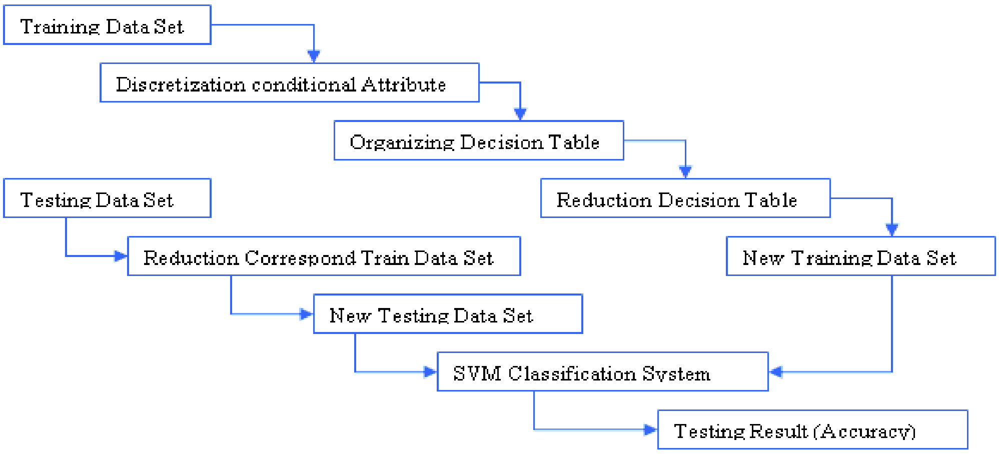

Training Data Set
Discretization conditional Attribute
Organizing Decision Table
Testing Data Set
Reduction Decision Table
Reduction Correspond Train Data Set
New Training Data Set
New Testing Data Set
SVM Classification System
Testing Result (Accuracy)

Fig. II Flow Diagram of A ROUGH-SVM using SVM & RSES

without affecting the system, are considered redundant. A minimal set of attributes after the redundancy removal is called reducts. Reduct preserves the degree of dependency and it cannot be reduced any further while still preserving the degree of dependency. Thus, reduct is a minimal set of feature attributes which has the same ability to discern groups as when the full set of feature attributes is used.

4. A Rough SVM Approach based on Support Vector Machine and Rough Sets Theory

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

- It is based on the original data only and does not need any external information, unlike probability in statistics or grade of membership in the Fuzzy set theory.
- It can reduce condition 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.

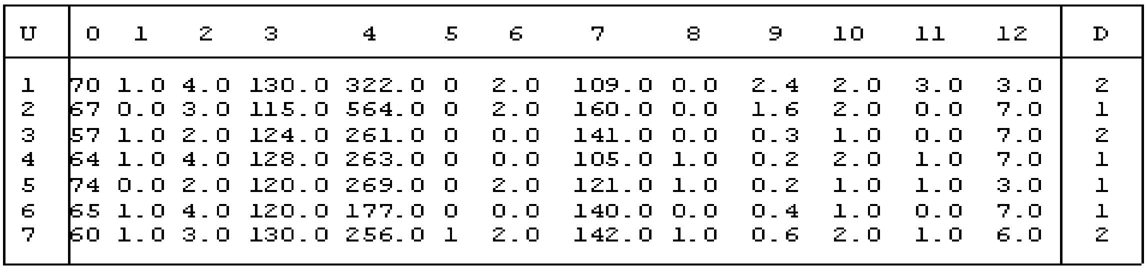
A kind of support vector machine classification system based on the Rough Sets pre-process is presented in this section. Given a training sample set, we firstly discrete them if the sample attribute values are continuous, and we can get a minimal feature subset that fully describes all concepts by attribute 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.

5. Results

Rough Set Exploration System (RSES version 2.2) and LIBSVM-2.85 has been used to carry out experiments on heart data set [5].

A. Heatr Data Set

In this data set, we have taken 270 observations, out of which 200 observations are taken for training and rest is for testing. it contains 13 (0- 12) attributes and one decision attribute (D). Table 1 shows some data samples of heart data, in which attribute "0" denotes age, attribute "1" denotes sex attribute "2" denotes pain type, attribute "3" denotes blood pressure, attribute "4" denotes serum cholesterol, attribute "5" denotes Blood sugar, attribute "6" denotes electrocardiograph, , attribute "7" denotes maximum heart rate, attribute "8" denotes angina, attribute "9" denotes old peak=ST, attribute "10" denotes slope ST, attribute "11" denotes no. of vessels, attribute "12" denotes thal. VD ={1,2}, where "1" denotes absence of heart disease and "2" denotes presence of heart disease. Table1 shows a sample of Heart Data Set.

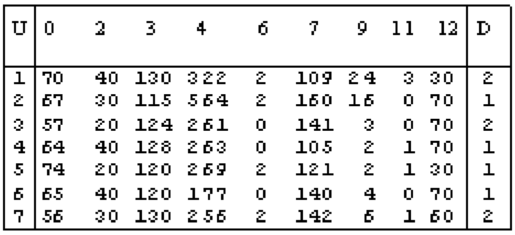


U	0	1	2	3	4	5	6	7	8	9	10	11	12	D
1	70	1.0	4.0	130.0	322.0	0	2.0	109.0	0.0	2.4	2.0	3.0	3.0	2
2	67	0.0	3.0	115.0	564.0	0	2.0	160.0	0.0	1.6	2.0	0.0	7.0	1
3	57	1.0	2.0	124.0	261.0	0	0.0	141.0	0.0	0.3	1.0	0.0	7.0	2
4	64	1.0	4.0	128.0	263.0	0	0.0	105.0	1.0	0.2	2.0	1.0	7.0	1
5	74	0.0	2.0	120.0	269.0	0	2.0	121.0	1.0	0.2	1.0	1.0	3.0	1
6	65	1.0	4.0	120.0	177.0	0	0.0	140.0	0.0	0.4	1.0	0.0	7.0	1
7	60	1.0	3.0	130.0	256.0	1	2.0	142.0	1.0	0.6	2.0	1.0	6.0	2

Table1 Sample of Heart Data Set

B. Discretization and Attributes Reduction of Heart Data Set

After discretization and attribute reduction 4 attributes (1, 5, 8, 10) out of 13 are found to be redundant. So remaining attributes (0, 2, 3, 4, 6, 7, 9, 11, 12) and their values are left respectively as shown in Table 2.

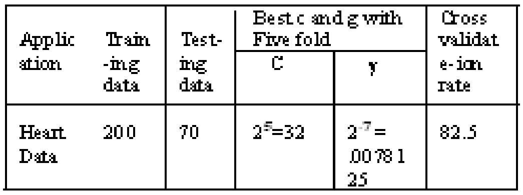


U	0	2	3	4	6	7	9	11	12	D
1	70	40	130	322	2	109	24	3	30	2
2	67	30	115	564	2	160	16	0	70	1
3	57	20	124	261	0	141	3	0	70	2
4	64	40	128	263	0	105	2	1	70	1
5	74	20	120	269	2	121	2	1	30	1
6	65	40	120	177	0	140	4	0	70	1
7	56	30	130	256	2	142	6	1	60	2

Table II Results After Discreztization and attribute Reduction

C. Apply SVM on Reduce Heart Data Set

The classifications experiments are conducted on reduced training and testing Heart data set. In these experiments, LIBSVM with RBF kernel function has been used. , RBF kernel parameters γ and the cost parameter C, has been determined using 5 – fold cross validation method as shown in Table III.

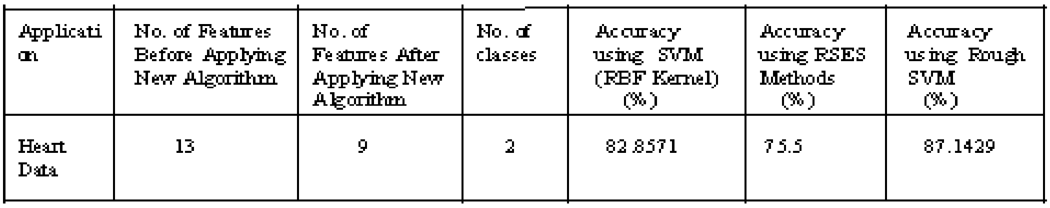


Application	Train-ing data	Test-ing data	Best c and g with Five fold		Cross validat e-ion rate
			C	γ	
Heart Data	200	70	2^5=32	2^{-7}= .0078125	82.5

Table III HEART DATA SET

6. Conclusion

In this paper, we have introduced a new data classification method: A Rough-SVM, which makes great use of the advantages of Support Vector Machine's greater generalization performance and Rough Set Theory in effectively dealing with vagueness and uncertainty information. Classification accuracy increased by 4. 29% shown in TableIV. So, we can observe that the classification accuracy using Rough-SVM is much better than general SVM and general RSES method.



Application	No. of Features Before Applying New Algorithm	No. of Features After Applying New Algorithm	No. of classes	Accuracy using SVM (RBF Kernel) (%)	Accuracy using RSES Methods (%)	Accuracy using Rough SVM (%)
Heart Data	13	9	2	82.8571	75.5	87.1429

Table.IV RESULTS USING DIFFERENT TECHNIQUE WITH HEART DATA SET

Acknowledgement

The authors would like to thank Mr. Chih-Jen Lin (Department of Computer Science National Taiwan University, Taipei 106, and Taiwan) for his useful software and many fruitful comments.

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 University, Taipei 106, Taiwan http://www.csie.ntu.edu.tw/~cjlin 2007

[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"Unlabeleddata classification via SVM and k-means Clustering". Proceeding of the International Conference

[7] RSES 2.2 User's Guide, Warsaw University http://logic.mimuw.edu.pl/~rses

[8] Z. Pawlak (1991): Rough sets: Theoretical aspects of reasoning about data. Dordrecht: Kluwer.

[9] Feng Honghai, Chen Guoshun, Wang Yufeng, Yang Bingru, Chen Yumei, "Rough Set Based Classification rules generation for SARS Patients". Proceedings of the 2005 IEEE, Engineering in Medicine and Biology 27th Annual Conference Shanghai, China, September 1-4, 2005

[10] R. S. Gautam, D. Singh, A. Mittal, "A Rough Set Classifilcation Based Approach to Detect Hotspots in NOAA/AVHRR Images", 2006 IEEE.

Received: October, 2009