DATA CLASSIFICATION USING SUPPORT VECTOR MACHINE

[1]DURGESH K. SRIVASTAVA, [2]LEKHA BHAMBHU

[1]Ass. Prof., Department of CSE/IT, BRCM CET, Bahal, Bhiwani, Haryana, India-127028

[2]Ass. Prof, Department of CSE/IT, BRCM CET, Bahal, Bhiwani, Haryana, India-127028

ABSTRACT

Classification is one of the most important tasks for different application such as text categorization, tone recognition, image classification, micro-array gene expression, proteins structure predictions, data Classification etc. 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. In this paper, a novel learning method, Support Vector Machine (SVM), is applied on different data (Diabetes data, Heart Data, Satellite Data and Shuttle data) which have two or multi class. 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 [1] & [2].

In our experiment, the support vectors, which are critical for classification, are obtained by learning from the training samples. In this paper we have shown the comparative results using different kernel functions for all data samples.

Keywords: *Classification, SVM, Kernel functions, Grid search.*

1. INTRODUCTION

The Support Vector Machine (SVM) was first proposed by Vapnik and has since attracted a high degree of interest in the machine learning research community [2]. Several recent studies have reported that the SVM (support vector machines) generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. Sims have been employed in 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. It has been shown that Sims is consistently superior to other supervised learning methods. However, for some datasets, the performance of SVM is very sensitive to how the cost parameter and kernel parameters are set. As a result, the user normally needs to conduct extensive cross validation in order to figure out the optimal parameter setting. This process is commonly referred to as model selection. One practical issue with model selection is that this process is very time consuming. We have experimented with a number of parameters associated with the use of the SVM algorithm that can impact the results. These parameters include choice of kernel functions, the standard deviation of the Gaussian kernel, relative weights associated with slack variables to account for the non-uniform distribution of labeled data, and the number of training examples.

For example, we have taken four different applications data set such as diabetes data, heart data and satellite data which all have different features, classes, number of training data and different number of testing data. These all data taken from RSES data set and http://www.ics.uci.edu/~mlearn/MLRepository.html [5]. 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 detail all experiments results. Finally, we have some conclusions and feature direction in Section 4.

2. SUPPORT VECTOR MACHINE

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

2.1 OVERVIEW OF SVM

SVMs are set of related supervised learning methods used for classification and regression [2]. They belong to a family of generalized linear classification. A special property of SVM is , SVM simultaneously minimize the empirical classification error and maximize the geometric margin. So SVM called Maximum Margin Classifiers. SVM is based on the Structural risk Minimization (SRM). SVM map input vector to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separate the data. The separating hyperplane is the hyperplane that maximize the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier will be [2].

We consider data points of the form

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

Where y_n=1 / -1 , a constant denoting the class to which that point xn belongs. n = number of sample. Each x $_n$ is p-dimensional real vector. The scaling is important to guard against variable (attributes) with larger varience. We can view this Training data , by means of the dividing (or seperating) hyperplane , which takes

$$w . x + b = o \qquad ----- (1)$$

Where b is scalar and w is p-dimensional Vector. The vector w points perpendicular to the separating hyperplane . Adding the offset parameter b allows us to increase the margin. Absent of b, the hyperplane is forsed to pass through the origin , restricting the solution. As we are interesting in the maximum margin , we are interested SVM and the parallel hyperplanes. Parallel hyperplanes can be described by equation

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

If the training data are linearly separable, we can select these hyperplanes so that there are no points between them and then try to maximize their distance. By geometry, We 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

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

This can be written as

$$y_i\ (\ w.\ x_i – b) \geq 1 \quad ,\ 1 \leq i \leq n \quad ------(2)$$

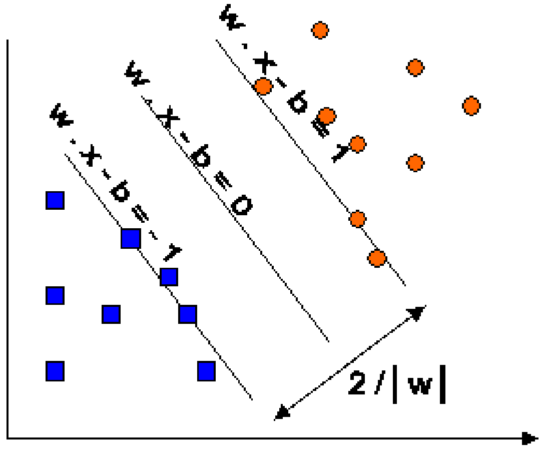

Figure.1 Maximum margin hyperplanes for a SVM trained with samples from two classes

Samples along the hyperplanes are called Support Vectors (SVs). A separating hyperplane with the largest margin defined by M = 2 / │w│ that is specifies support vectors means training data points closets to it. Which satisfy?

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

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

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

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

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

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

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

$$= 1/2 \, w^T w - \sum_{i=1}^{l} \alpha_i \; (y_i (w^T x_i + b) - 1) \text{ ---(5)}$$

Where α_i is a Lagranges multiplier .The search for an optimal saddle points (w_0, b_0, α_0) is necessary because Lagranges must be minimized with respect to w and b and has to be maximized with respect to nonnegative αi ($\alpha_i \geq 0$). This problem can be solved either in primal form (which is the form of w & b) or in a dual form (which is 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) with respect to saddle points (w_0, b_0, α_0).

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

$$\text{i .e} \qquad w_0 = \sum_{i=1}^{l} \alpha_i \; y_i \; x_i \qquad \text{-----------(6)}$$

$$\text{And} \qquad \partial L / \partial b_0 = 0$$

$$\text{i .e} \qquad \sum_{i=1}^{l} \alpha_i \; y_i = 0 \qquad \text{-----------(7)}$$

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

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

In order to find the optimal hyperplane, a dual lagrangian (L_d) has to be maximized with respect to nonnegative α_i (i .e. α_i must be in the nonnegative quadrant) and with respect to the equality constraints as follow

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

$$\sum_{i=1}^{l} \alpha_i \; y_i = 0$$

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].

2.2 KERNEL SELECTION OF SVM

Training vectors x_i are mapped into a higher (may be infinite) dimensional space by the function Φ. Then SVM finds a linear separating hyperplane with the maximal margin in this higher dimension space .C > 0 is the penality 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 : $K (x_i , x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad , \quad \gamma > 0$
- Sigmoid kernel: $K (x_i , x_j) = \tanh(\gamma x_i^T x_j + r)$

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

1. The RBF kernel nonlinearly maps samples into a higher dimensional space unlike to linear kernel.
2. The RBF kernel has less hyperparameters than the polynomial kernel.
3. The RBF kernel has less numerical difficulties.

2.3 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, 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 dataset and then get the classifier. After that, use the classifier to classify the testing dataset to get the generalization accuracy.

3. INTRODUCTION OF ROUGH SET

Rough set is a new mathematic tool to deal with un-integrality and uncertain knowledge. It can effectively .analyze and deal with all kinds of fuzzy, conflicting and incomplete information, and finds out the connotative knowledge from it, and reveals its underlying rules. It was first put forward by Z.Pawlak, a Polish mathematician, in 1982. In recent years, rough set theory is widely emphasized for the application in the fields of data mining and artificial intelligence.

3.1 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

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

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 $Ui \cap Uj = \Phi$, i ,j = 1to m and i$\neq$ 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

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

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

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 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 $C \subset Q$, $D \subset Q$, $C \cap D = \Phi$. 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 γc (D) and is defined by

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

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

$POS_C\ (D)$ 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.

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

An attribute a is 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].

4 RESULTS OF EXPERIMENTS

The classification experiments are conducted on different data like Heart data, Diabetes data, Satellite data and Shuttle data. These data taken from http://www.ics.uci.edu/~mlearn/MLRepository.html and RSES data sets . In these experiments, we done both method on different data set. Firstly, Use LIBSVM with different kernel linear , polinomial , 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. All three data sets, diabetes , heart, and satellite, are 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. 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 represent 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 all data to predict the accuracy in seconds.

Applicat-ions	Training data	Testing data	Best c and g with five fold		Cross validation rate
			C	γ	
Diabetes data	500	200	2^{11}=2048	2^{-7}= .0078125	75.6
Heart Data	200	70	2^{5}=32	2^{-7} = .0078125	82.5
Satellite Data	4435	2000	2^{1}=2	2^{1}=2	91.725
Shuttle Data	43500	14435	2^{15}= 32768	2^{1}=2	99.92

Table 1

Applications	Total Execution Time to Predict	
	SVM	RSES
Heart data	71	14
Diabetes data	22	7. 5
Satellite data	74749	85
Shuttle Data	252132.1	220

Table 2: Execution Time in Seconds using SVM & RSES

Fig. 2, 3 shows, Accuracy comparison of Diabetes data Set after taking different training set and all testing set for both technique (SVM & RSES) using RBF kernel function for SVM and Rule Base Classifier for RSES.

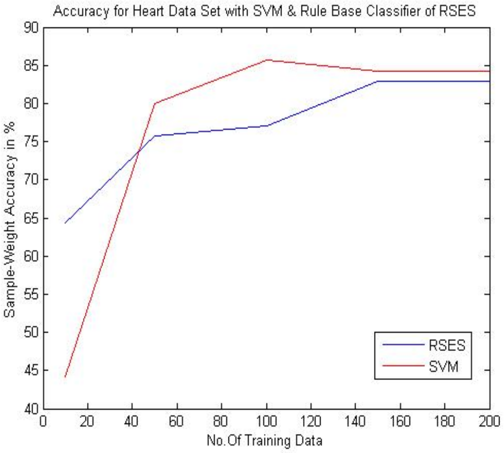

Fig :2 Accuracy of Heart data with SVM & RSES

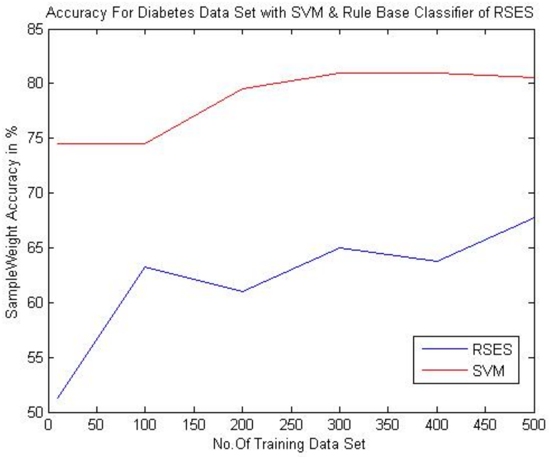

Fig: 3 Accuracy of Diabetes data with SVM & RSES

Applications	Training data	Testing data	Feature	No. Of Classes	Using SVM (with RBF kernel)	Using RSES with Different classifier			
						Rule Based Classifier	Rule Based Classifier with Discretization	K-NN Classifier	LTF Classifier
Heart data	200	70	13	2	82.8571	82.9	81.4	75.7	44.3
Diabetes data	500	200	8	2	80.5	67.8	67.5	70.0	78.0
Satellite data	4435	2000	36	7	91.8	87.5	89.43	90.4	89.7
Shuttle Data	43500	14435	9	7	99.9241	94.5	97.43	94.3	99.8

Table 3: Compare with Rough Set Classifiers

5 CONCLUSION

In this paper, we have shown the comparative results using different kernel functions. Fig 2 and 3 shows the comparative results of different data samples using different kernels linear, polynomial, sigmoid and RBF. The experiment results are encouraging .It can be seen that the choice of kernel function and best value of parameters for particular kernel is critical for a given amount of data. Fig 3 shows that the best kernel is RBF for infinite data and multi class.

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” . Deptt of Computer Sci. National Taiwan Uni, 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. 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 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 http://logic.mimuw.edu.pl/»rses ,January 19, 2005

[9] Eva Kovacs, Losif Ignat, “Reduct Equivalent Rule Induction Based On Rough Set Theory”, Technical University ofCluj-Napoca.

[9] RSES Home page http://logic.mimuw.edu.pl/»rses

BIOGRAPHY:

Mr Durgesh K. Sriavastava received the degree in Information & Technology (IT) from MIET, Meerut, UP, INDIA in 2006. He was a research student of Birla Institute of Technology (BIT), Mesra, Ranchi, Jharkhand, INDIA) in 2008. Currently, he is an Assistant Professor (AP) at BRCM CET, Bahal, Bhiwani, Haryana, INDIA. His interests are in Software engineering & modeling and design, Machine Learning.

Mrs Lekha Bhambhu received the degree in Computer Science & Engineering from BRCM CET, Bahal, Bhiwani, Haryana, INDIA. she was a research student of CDLU, Sirsa, Haryana, INDIA. Currently, she is an Assistant Professor (AP) at BRCM CET, Bahal, Bhiwani, Haryana, INDIA. Her interests are in Operating System, Software engineering.