Data Classification with Support Vector Machine Kernel Function

Serri Ismael Hamad

Department of Computer Science, College of Education for Pure Sciences, University of Thi-Qar, Iraq
serriismael@utq.edu.iq

ABSTRACT

Classification,is one of the most, important tasks for ,variousa pplication such,as,data ,Classification ,image, classification ,text ,categorization, micro- array, gene, expression, tone ,recognition, proteins, structure predictions,etc.The majority, of today's supervised, classification, algorithms ,are based on traditional, statistics, which can produce optimal, results when the sample, size approaches, infinity." In practice, however, only finite ,samples may be obtained. In this paper, a unique learning, method called ,Support Vector, Machine (SVM) is used, to data with two or more, classes such as Diabetes, data, Satellite, data , Shuttle data, and Heart, data. SVM is a strong ,machine learning ,algorithm that has achieved ,substantial success in, a variety ,of fields.They were first introduced, in the early 1990s, and they ,sparked ,a surge, in interestt, in machine ,learning." Vapnik laid, the groundwork, for SVMs, which are gaining ,traction in the field of machine learning, because to their many appealing features, and promising, empirical ,outcomes."SVM method does not suffer,the limitations, of data,dimensionality, and limited samples,[1] and [2]". The SVM which are ,important for classification, are learned from, the training ,data in our experiment.For all data samples, we have given,comparison findings, using different, kernel functions, in this research."

Keywords: Classification,SVM,Kernelfunctions,model selection

1. Introduction

The Support Vector Machine is one of the classical machine learning techniques that can still help solve big data classification problems . Vapnik was the first, to suggest SVM and it has since piqued the interest of the machine learning research ,community [2]. Data classification tone, recognition, image classification, and object detection microarray gene expression data analysis. Sims has been found to outperform other supervised learning algorithms on a constant basis. However the, performance of SVM is highly dependent, on how the cost parameter and kernel parameters are chosen for some datasets. As a result in order to determine the best parameter value the user usually needs to undertake significant, cross validation. Model selection is the term used to describe this procedure. We have experimented with a number of factors linked with the use of the SVM algorithm that can affect the findings we have experimented with a number of parameters related with the use of the SVM algorithm that can impact the results. The number of training examples as well as the 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. For example, we've chosen four different application data sets such as diabetes heart and satellite data each of which has its own set of features classes training data, and testing data. These are all data from the RSES data set and http://www.ics.uci.edu/~mlearn/MLRepository.htm [5].
The following is a breakdown, of the paper's structure. In the next section we'll go over some background information such as some basic SVM ideas, kernel function selection and so on as well as SVM model selection (parameter selection). All of the, outcomes of the experiments are detailed in Section 3. Lastly Section 4 contains some findings and feature direction.

2. Basic Concepts Of Support Vector Machine

We'll go over some basic SVM ideas different kernel functions and SVM model selection (parameters selection) in this part.

2.1- Overview Of The Support Vector Machine

The SVMs are a collection of supervised learning algorithms 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 andmaximize 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 hyprplanes 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 separating) hyperplane which takes

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

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 theparallel 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

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

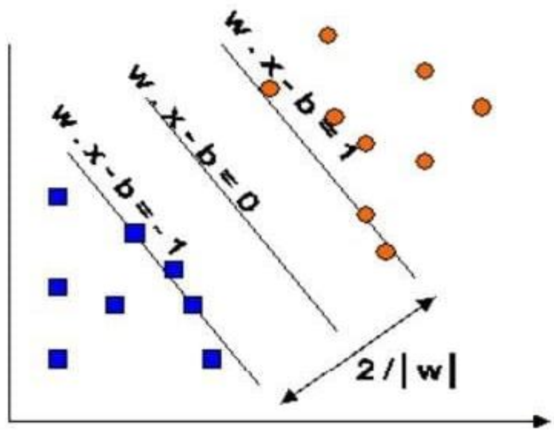

Figure (1): Hyperplanes of maximum margin for an SVM trained with samples from two classes

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?

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

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

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

The number of training data points is, denoted by the letter l. A learning machine should minimize $\|w\|^2$while considering inequality restrictions in order to Identify the optimal separation hyperplane with a maximul margin.

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

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 αi 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 α_i ($\alpha_i \geq 0$). This problem can be solved either in primal form (which is the form of w and 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).

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

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

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

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

Equations (6) and (7) are substituted in equation (5). The primal form is transformed into a dual form.

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

A dual lagrangian (L_d) must be maximized with regard to nonnegative I (i.e. I must be in the nonnegative quadrant) and the equality constraints as follows in order to discover the best hyperplane.

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

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

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]and[2] .

2.2- KERNEL SELECTION OF SVM:

The function Φ Maps training vectors xi onto a higher (perhaps infinite) dimensional space**.** Then in this higher-dimensional space SVM determines a linear separating hyperplane with the maximum margin where C > 0 is the error term's penality parameter.

Furthermore (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 best kernel function is also a research topic.

However for general purposes there are some popular kernel functions [2] and [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, using this parameter set, apply it to the training dataset to obtain a classifier. Then, to acquire the generalization accuracy, apply the classifier to categorize the testing dataset.

3. INTRODUCTION OF ROUGH SET

The Rough set is a novel mathematical tool for dealing with non-integrality and ambiguous knowledge. It can effectively assess and deal with a wide range of ambiguous, conflicting, and incomplete data, extracting connotative knowledge and revealing underlying rules. It was first proposed in 1982 by Z. Pawlak a Polish mathematician. In recent years, rough set theory has received a lot of attention for its use in data mining and artificial intelligence.

3.1 THE BASIC DEFINITIONS OF ROUGH SET

Assume S is a four-element information system.

$S = (U, Q, V, f)$ 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 ⊆ 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 > ∈ 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 Ui ∈ 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 ⊆ U and P inferior approximation P1 and superior approximation P1 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 which is based on discovering a reduct from the original set of qualities is successfully employed in feature selection. The initial set of attributes will not be used by data mining methods,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 people who make decisions Classes of equivalency 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 γc (D) = 1 then, C determines D functionally. Data set U is called consistent if γc (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 POSR (D) = POSC(D) and R has no R' subset $R' \subset R$ so that POSR'.(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. There could be numerous reducts for the set of qualities C. The core of C is the set of qualities that belong to the intersection of all reducts of C set.

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

An attribute a is indispensable for C if POS_C (D) $\neq$ $POS_{C[a]}$ (D). C's core consists of the union of all of the language's essential features. There are two definitions for the core More detailed information on RSES can be found in. [1]and[2].

4. Results of Experiments

In the classification, studies various sorts of data are employed including heart data, diabetes data, satellite data and shuttle data. These data taken

From http://www.ics.uci.edu/~mlearn/MLRepository.htm l and RSES data sets. We tested both methods on different data sets in these trials. To begin use LIBSVM with several kernels such as linear, polinomial and sigmoid and RBF[5]. RBF kernel is employed. As a result two parameters must be set: the RBF kernel parameter and the cost parameter C.Table(1) lists the three datasets utilized in the trials major characteristics All three data sets (diabetes, heart, and satellite) have been combined 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 the RSES Tool set is used to classify all data sets using various classifier techniques such as Rule Based Classifiers Rule Based Classifiers with Discretization K-NN classifier and LTF (Local Transfer Function) Classifier. The hardware ,platform used in the experiments is a workstation with Pentium-IV1GHz ,CPU 256MB RAM and the Windows XP (using MS-DOS Prompt).The findings of the various experiments are represented in the following three tables.

Table(1) displays the optimum result for various RBF parameter values (C) and cross validation rate using the grid search method [5] and [6]. Table(2) displays the Total execution time for all data to predict the accuracy in seconds.

Applic at-ions	Train ing data	Testi ng data	Best c and g with five fold		Cross validati on rate
			C	γ	
Diabet es data	500	200	2^{11}=2048	2^{-7}= .0078125	75.6
Heart Data	200	70	2^{5}=32	2^{-7} = .0078125	82.5
Satellit e 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): displays the best value of different RBF parameter

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): Time to Execution in Seconds using SVM and RSES

Figure (2, 3) displays Diabetes data accuracy comparison Set utilizing RBF kernel function for SVM and Rule Base Classifier for RSES after taking distinct training and testing sets for both techniques (SVM and RSES).

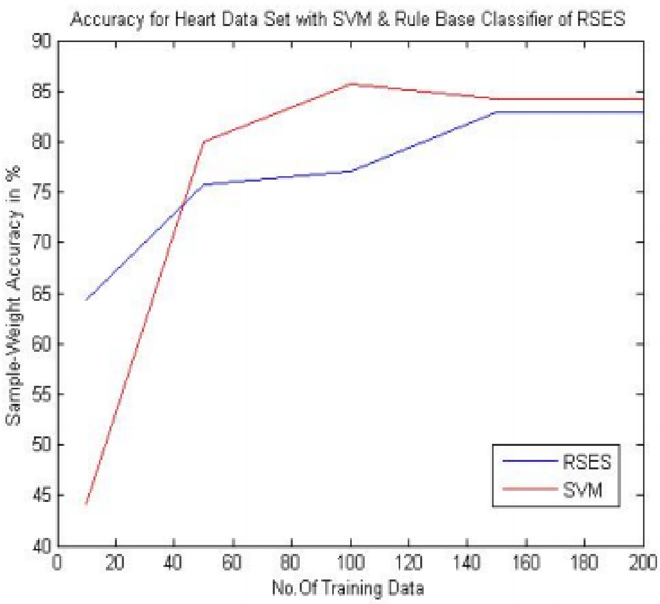

Figure (2): SVM and RSES improves the accuracy of heart data.

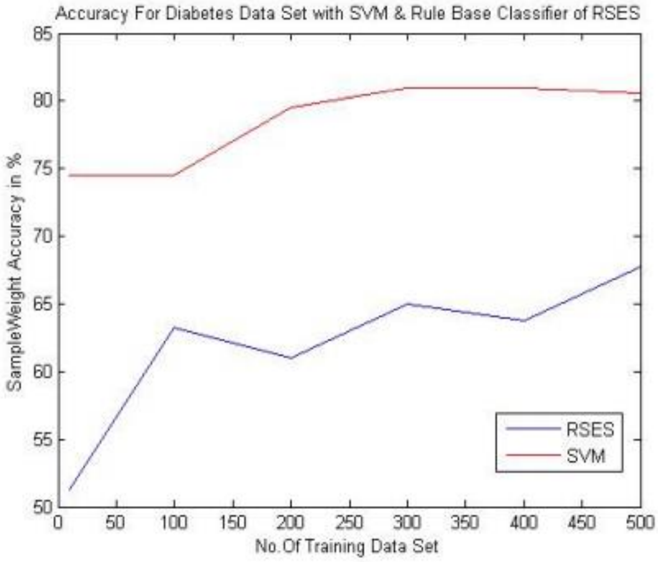

Figure (3): Diabetes data accuracy with SVM and RSES

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

5 – Conclusion

We have given comparison findings utilizing several kernel functions in this research. Figures (2) and (3) displays the outcomes of several data samples utilizing various kernels such as linear polynomial, sigmoid and RBF. The outcomes of the experiment are favorable. It can be observed that for a given amount of data, the kernel function and optimum parameter values for ,that kernel are crucial. Figure (3) displays that RBF is the optimal kernel for infinite data and multi-class problems.

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