Solving Classification Problem using Reduced Dimension and Eigen Structure in RSVM

Meeta Pal Deepshikha Bhati Baijnath Kaushik Haider Banka

ABSTRACT

Support vector machine (SVM) is a recent method to classify the data. SVM has been proved as a powerful tool for solving classification problem. The problem with complex dataset incurs significant complexity while classifying and its efficiency also cost very much. We propose a reduced set support vector machine based on Eigen structure, to classify dataset having multiple features. In this paper, Eigen vectors use to present the whole data in reduced dimensions. This minimize the task of classification by propose method and cost is reduced while efficiency is improved with the increase complexity of data. The proposed method takes a random chunk of data followed by Eigen structure use to reduce the dimension of the data. So as classification problem solve efficiently. We have compared the proposed method with SVM and RSVM. The result signifies that the proposed method gives better result in comparison to SVM and RSVM.

Keywords

RSVM, SVM, Support Vector based on Eigen Structure.

1. INTRODUCTION

Support vector machine (SVM) is a statistical technique used to classify the data. It is a supervised method which used training data to identify a model. This model contains the support vectors. A support vector machine construct a hyper plane or set of hyper plane in a high-or-infinite dimensional space which can be used for classification, regression or other tasks, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data point of any class, since in general the larger the margin the lower the generalization error of the classifier.

The original SVM algorithm was invented by Vladimir N. Vapnik and Alexey Ya.Chervonenkis in 1963. The current standard incarnation (soft margin) was proposed by corning Cortes and Vapnik in 1993 and published in 1995. [1]

SVM is mostly used in field of machine learning and pattern recognition. It is based on structural risk minimization. That means it searches for a decision boundary, which aims at providing a trade-off between hypothesis space complexity and quality of fitting the training data. SVM finds hyper-plane for linearly separable data as well as for linearly not separable (non linear) data. This hyper plane gives the largest minimum distance (margin) to the training example. For non linear data it uses different types of kernel function like Gaussian Radial basis function, multilayer perceptron, Quadratic kernel, polynomial kernel etc.

The SVM has been applied for different applications like pattern recognition, geometric camera calibration, text categorization, handwritten character recognition, face recognition and in classification of object into discrete classes.

In next section we will see Reduced Support Vector Machine (RSVM).

1.1 RSVM

Recently the Reduced Support Vector Machine (RSVM) was proposed as an alternate of the standard SVM. Motivated by resolving the difficulty of handling large data sets using SVM with non-linear kernels, it preselects a subset of data as support vectors and solves a smaller optimization problem.

The procedure has two steps:

a. It finds a sample/ template data

b. Applying SVM on selected Template

In order to deal with large data sets, the reduced support vector machine (RSVM) was proposed for the practical objective to overcome some computational difficulties as well as to reduce the model complexity.

Basically RSVM is used from the view point of sampling design, its robustness, and the spectral analysis of the reduced kernel [2].

Our main concern is to reduce the size of training set as well as to increase the efficiency.

In next section we have describe the RSVM algorithm given by Su-Yun Huang and Yhu-Jye Lee [2].

1.2 RSVM Formulation

Consider the problem of classifying points into two classes, A- and A+. We are given a training data set $\{(x^i y_i)\}_{i=1}^{m}$, where is an input vector and $y_i \epsilon \{-1, 1\}$ is a class label, indicating one of the two classes A- and A+, to which the input point belongs. We represent these data points by an m×n matrix A, the ith row Ai corresponds to the ith input data point. We use alternately A_i (a row vector) and x^i (a column vector) for the same ith data point depending on the convenience. The main goal of the classification problem is to find a classifier that can predict correctly the unseen class labels for new data inputs. It can be achieved by constructing a linear or nonlinear separating surface f(x)=0, which is implicitly defined by a kernel function. We classify a test point x to A+ if f(x) ≥0, otherwise, to A-. In conventional SVM as well as many kernel-based learning algorithms [3]-[5], generating a nonlinear separating surface has to deal with a fully dense kernel matrix with the size of the number of training examples. When training a nonlinear SVM on a massive data set, the huge and dense full kernel matrix will lead to some computational difficulties as follows:

a. The size of a mathematical programming problem.

b. Dependency of the nonlinear separating surface on most of the data set, which creates unwieldy storage problems that hinder the use of nonlinear kernels for massive data sets.

To avoid these difficulties and to cut down model complexity, the RSVM uses a very small random subset of size m̃, where m̃<<m, for building up the separating surface which plays a similar role of support vectors. We denote this random subset by Ã, which is used to generate a much smaller rectangular matrix K (A, Ã`) $\epsilon$ $R^{m\times \tilde{m}}$. The reduced kernel matrix K (A, A`), is to cut problem size, computing time, and memory usage as well as to simplify the characterization of the nonlinear separating surface.

We now briefly describe the RSVM formulation for binary classification, which is derived from the generalized support vector machine (GSVM) [3] and the SSVM [6]. The RSVM starts from a standard two-norm soft margin SVM, and next it appends the term $\gamma^2/2$to the objective function to be minimized and result in the following minimization problem:

$$\min_{(u,\gamma,\epsilon)} \frac{C}{2}\|\epsilon\|_2^2 + \frac{1}{2}(\|u\|_2^2 + \gamma^2)$$

$$D\{K(A,A')u - 1\gamma\} \geq 1 - \varepsilon \qquad \varepsilon \geq 0$$

Where, C is a positive number, for balancing training error and regularization term in the objective function. We call it weight parameter. We note that the nonnegative constraint $\epsilon \geq 0$ can be removed because of the term $\| \varepsilon \|_2^2$ in the objective function. We use an m×m diagonal matrix D, where $D_{ii}=y_i \in \{-1, 1\}$, to specify the corresponding class membership of each input point. If we let v=Du, then $\| v \|_2^2 = \| u \|_2^2$. Thus, the previous problem is equivalent to

$$\min_{(u,\gamma,\epsilon)} \frac{C}{2}\|\epsilon\|_2^2 + \frac{1}{2}(\|v\|_2^2 + \gamma^2) \qquad (1)$$

$$D\{K(A,A')v - 1\gamma\} \geq 1 - \varepsilon \qquad (2)$$

At a solution ε takes the form $\varepsilon = (1-D\,\{K\,(A, A`)\,v-1\gamma\})_+$. Next, we convert the problem given by (1) and (2) into an equivalent SVM, which is an unconstrained optimization problem as follows:

$$\min_{(v,\gamma)\epsilon R^{m+1}} \frac{C}{2} \| (1 - D\{\mathrm{K}(\mathrm{A},\mathrm{A`})\mathrm{v} - 1\gamma\}) + \|_2^2 + \frac{1}{2}(\| v \|_2^2 + \gamma^2) \qquad (3)$$

Instead of using the full kernel matrix K (A, A`), we replace it with a reduced kernel matrix K (A, A`), where Ã consist of m̃ random columns from A, and the problem becomes

$$\min_{(\tilde{v},\gamma)\epsilon R^{m+1}} \frac{C}{2} \| (1 - D\{\mathrm{K}(\mathrm{A},\mathrm{A`})\tilde{\mathrm{v}} - 1\gamma\}) + \|_2^2 + \frac{1}{2}(\| \tilde{\mathrm{v}} \|_2^2 + \gamma^2) \qquad (4)$$

In solving the RSVM (4), a smooth approximation p (z, α) to the plus function is used [6]. The p function defined below can provide a very accurate approximation. The RSVM then solves the following approximation unconstrained minimization problem for a general kernel K (A, Ã`):

$$\min_{(\tilde{v},\gamma)\epsilon R^{m+1}} \frac{C}{2} \| \left(1 - D\{\mathrm{K}(\mathrm{A},\mathrm{A`})\tilde{\mathrm{v}} - 1\gamma\}, \alpha\right) + \|_2^2 + \frac{1}{2}(\| \tilde{\mathrm{v}} \|_2^2 + \gamma^2) \qquad (5)$$

Where p (z, α) is defined component wise by

The function p (z, α) converges to $(z)_+$ as α goes to infinity. Since the RSVM has already reduced the model complexity via using a much smaller rectangular kernel matrix (corresponding to using less support vectors in constructing the decision boundary), we will suggest to use a larger weight parameter C in the RSVM than in a conventional SVM. The solution to the minimization problem (4) or (5) leads to a nonlinear separating surface of the form

$$\sum_{i=1}^{\tilde{m}} \tilde{\mathrm{v}}\mathrm{k}(\tilde{A}_i, x) - \gamma = 0$$

In fact , the reduced set Ã is not necessarily to be a subset of training set [8]. The minimization problem (5) retains the strong convexity and differentiability properties in the space for any arbitrary rectangular kernel. Hence, we can apply the Newton-Armijo method [6] directly to solve (5). The existence and uniqueness of the optimal solution are also guaranteed. Moreover, the computational complexity of solving problem (5) by the Newton -Armijo method is O (m̃) while solving the nonlinear SSVM with the full square kernel is O (m^3) [6]. Typically, m̃<< m. the numerical test in [8] on the adult data set[9] shows that sample standard deviation of test set correctness for 50 replicate runs using $\tilde{A} \in R^{326\times123}$ out of $\tilde{A} \in R^{32562\times123}$ is less 0.001. The smallness of the standard error provides guidance for determining m̃.

In summary, the RSVM can be split into two parts. First, it select a small random subset $\{K(\tilde{A}_{1,}\cdot),\{K(\tilde{A}_{2,}\cdot),\cdots,\{K(\tilde{A}_{\tilde{m},}\cdot)\}$ from the full-data bases $\{k(A_i,.)\}_{i=1}^m$ for building the separating surface prior to training, while the conventional SVMs use a set of support vectors which are determined after training for building the surface. When projected onto the separating surface, the full-data bases are likely highly correlated with possibly heavy overlaps, which make room for model reduction. Second, the RSVM determines the best coefficients of the selected kernel functions by solving the unconstrained minimization problem (4) or (5) using the entire data though the RSVM uses only a small portion of kernel bases, it can still keep most of the relevant pattern information given by the entire training set.

2. PROPOSED METHOD

In the proposed method, we have applied the SVM algorithm on the Eigen structure on the whole dataset. The main idea of our algorithm is of two steps broadly:

1) Find the compressed Eigen structure of dataset:

The process of compressing the dataset is as follows:

a) Let's the dataset is, if

$$X = \begin{bmatrix} x_1 & \cdots & x_m \\ \vdots & \ddots & \vdots \\ x_n & \cdots & x_{mn} \end{bmatrix}_{m\times n}$$

b) Subtract the mean

$$X_{ij} = X_{ij} - \frac{\sum_{j=1}^{m} X_{ij}}{m}$$

And say after subtracting the mean, new matrix is N.

c) Find the covariance matrix, such that element

$$X_{ij} = cov\left(X_{ij}, X_{ji}\right)$$

Where $cov(X,Y) = \frac{\sum_{i=1}^{m}(X_i - \mathbb{X})(Y_i - \mathbb{Y})}{m-1}$

$\mathbb{X} = mean\ value\ of\ X$

And

$\mathbb{Y} = mean\ value\ of\ Y$

In a covariance matrix,

$$X_{ij} = X_{ji} = cov(X_{ij}, X_{ji}) = \sum \frac{(X_{ij} - \mathbb{X})(X_{ji} - \mathbb{X})}{m-1}$$

d) Find Eigen vectors and Eigen values of covariance matrix. And sort Eigen vector corresponding to their Eigen value in a decreasing order.

e) Select Eigen vectors as selected-vector matrix Ƒ. You can reduce no of dimension here i.e. if you want to reduce data 50%, and then select only 50% of Eigen vectors.

f) Then find finally reduced data R as follows:

$$R = \text{Ƒ}^t \times N^t$$

2) Finding the support vector

In second step one can classify the dataset by finding hyper plane. Take the reduced data R, find a hyper plane, set of points x, satisfying,

$$wx + b = 0$$

Where w is normal vector to the hyper plane.

Figure 1 shows the linear separation of data. And It just classify data into two classes.

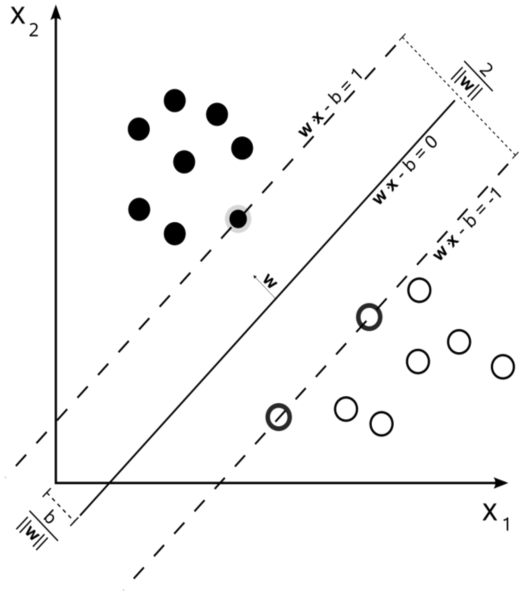

Fig. 1.Linear separation of data using hyper plane

The parameter $\frac{b}{\|\mathbf{w}\|}$ determine the offset of hyperplane from the origin along the normal vector w. For nonlinear classification, one can use kernel functions i.e. Gaussian Radial basis function, multilayer preceptron, Quadratic kernel, polynomial kernel etc.

3. RESULT ANALYSIS

We have tested our method on BUPA, PIMA, and PHONEME dataset[+]. BUPA dataset is created by a company which provides private medical insurance, named as The **B**ritish **U**nited **P**rovident **A**ssociation Limited. This dataset contains six attributes and can be classified into two classes. PIMA dataset is of Indians Diabetes dataset. This dataset contains seven attributes and it can also be classified into two classes. PHONEME dataset is basically used to classify nasal and oral sounds. This dataset contains five attributes and can be classified into two classes. We have tested SVM (Simple Support Vector Machine), RSVM (Reduced Support Vector Machine) and Our Method using ten-fold dataset.

We have taken training and testing data size for each dataset as following:

Dataset	BUPA	PIMA	Phoneme
Training data	310x6	200x	541
Testing data	35x6	25	50

Table1. Training and testing data size for each dataset

The comparison analysis of number of support vector using different method is as following:

Method	No. of Support Vector		
	BUPA	PIMA	Phoneme
SVM	256	174	259
RSVM	143	95	153
Our Method	260	137	246

Table2. Comparison analysis of generated no. of support vector using different method

Comparison analysis based on Accuracy is as follows:

Table3. Comparison analysis based on Accuracy

Method	BUPA	PIMA	Phoneme
SVM	57.14	60 %	78 %
RSVM	54.28	48 %	78 %
Our Method	57.14	72 %	80 %

Comparison Analysis Based on Time Taken is as follows:

Table4. Comparison Analysis Based on Time Taken

Method	BUPA	PIMA	Phoneme
SVM	1.0776	0.3182	0.5396
RSVM	0.034	0.0302	0.0509
Our Method	0.0774	0.0381	0.0846

Our method is used to classify data after reducing its dimensions so that its space and time complexity can be reduced.

We have tested SVM, RSVM and Our method on three different test datasets. Following graph shows accuracy of these methods tested on BUPA dataset.

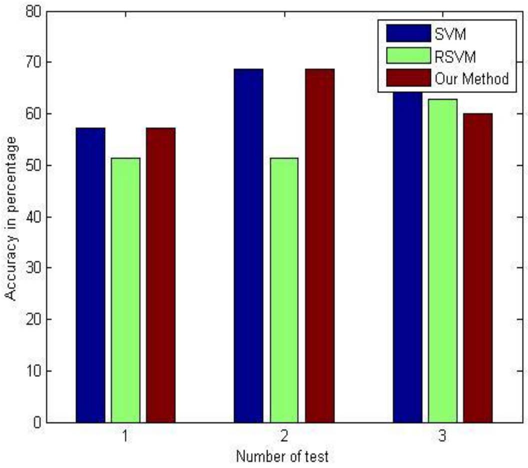

Fig.2. Accuracy of BUPA dataset

Following graph shows accuracy of SVM, RSVM and Our method tested on three test dataset of PIMA.

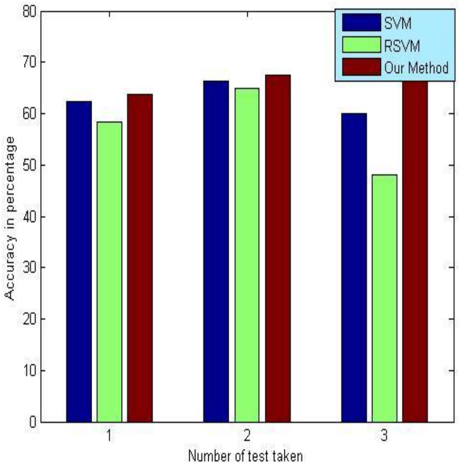

Fig. 3.Accuracy of PIMA dataset

Following graph shows accuracy of SVM, RSVM and Our method tested on three test dataset of PHONEME.

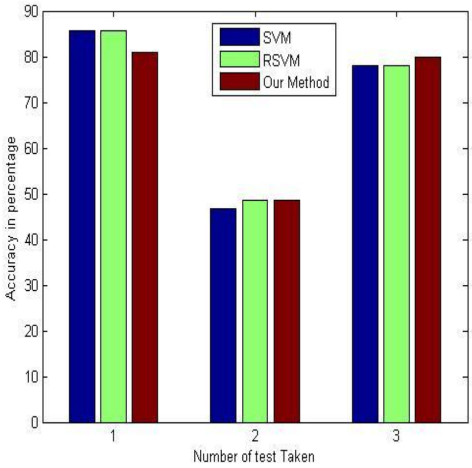

Fig. 4.Accuracy of PHONEME dataset

An analysis has been done for showing the actual result. We have taken different dataset and plotted a graph by taking test number (i.e. 1st test, 2nd test etc.) on X-axis and their corresponding accuracy on Y-axis. This graph shows that our method performs better than RSVM.

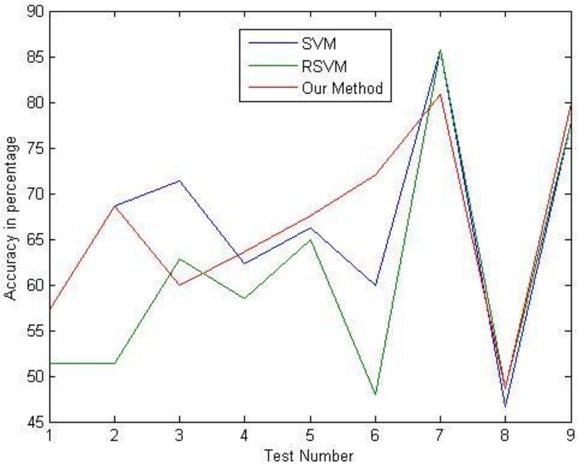

Fig. 4.Spectral result analysis

4. CONCLUSION AND FUTURE WORK

For large data set, Numbers of Support Vector are usually large. Because of that the computational complexity increases and the training time as well as testing time also increases. To simplify the SVM and reducing the number of Support Vectors, RSVM works well. But most of the time it reduces the testing accuracy. As results shows that the proposed method, not only reduce the training time but also improves the testing accuracy. Although time taken by our algorithm is more than that of RSVM, but always less than SVM.

5. REFERENCES

[1] Cortes.C. ; Vapnik, V. (1995). "Support-Vector Networks." Machine Learning 20(3) :273. Doi :10.1007/BF00994018.

[2] Yuh-aye Lee and Su-Yun Huang, "Reduced Support Vector Machine: A Statistical Theory," IEEE Transaction on Nueral Networks. Vol. 18.No.1, January 2007.

[3] V.N. Vapnik, The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.

[4] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in Advances in Nueral Information Processing Systems 9, M.C. Mozer, M.I. Jordan, and T. Petsche, Eds. Cambridge, MA: MIT Press,1997. Pp. 155-161

[5] O.L.Mangesarian and D. R. Musicant, "Large Scale kernel regression via linear Programming" Data Mining Inst., Comp. Sci. Dept., Univ. Wisconsin, Madison, WI, Tech. Rep. 99-02, Aug. 1999[Online]. Available: ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/99-02.ps.

[6] D. P. Wiens. Designs for approximately linear regression: Two optimality properties of uniform design. Statist. & probab. Letters, 12: 217-221, 1991

[7] D. P. Wiens. Minimax Designs for Approximately Linear Regression. J. Statist. Plann, inference, 31:353-371, 1992

[8] Y. J. Lee and O. L. Mangesarian. SSVM: A Smooth Support Vector Machine. Computational Optimization and Applictions, 20: 5-22, 2001.

[9] A. Smola and B. Scholkopf. Sparse Greedy Matrix Approximation for Machine Learning. In Proc. 17th international conf. on Machine Learning, Pages 911-918 Morgan Kanfmam, San Fransisco, CA,2000.

[10] Chistopher J. C. Burges,"A tutorial on Support Vector Machines for Pattern Recognition, " © Kluwer Acadimics Publishers, Boston. Manufactured in the Neitherland. Apeared in: Data Mining and Knowledge Discovery 2, 121-167 1998.

[11] J.Platt, "Fast training of SVMs using sequential minimal optimization", In B. Sch¨olkopf, C.Burges and A.Smola (ed.), Advances in Kernel Methods: Support Vector Learning, MIT Press, Cambridge, MA, 1999, 185-208.

[12] G. Fung and L.Mangasarian, "Proximal support vector machine classifiers", Proceedings of the 7th ACM conference on knowledge discovery and data mining, ACM, 2001, 77-86.