Available online at www.sciencedirect.com

ScienceDirect

Materials Today: Proceedings XX (2016) XXX–XXX

materialstoday: PROCEEDINGS

www.materialstoday.com/proceedings

Efficient Decision Tree Based Data Selection and Support Vector Machine Classification

P.Arumugam, P.Jose*

Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli, 628012, Tamilnadu, India

Abstract

Today' real world data bases witnessed significant increase in the amount of data in digital format, due to the widespread use of datasets and storage system. There is a need to developing fast and highly accurate algorithms to automatically classify large data. It becomes a vital part of the machine learning and knowledge discovery. The main intention of this paper is however data sizes increases, our proposed method make faster computation and scalable machine learning algorithm is used to learn faster from the labelled training data. Due to its strong mathematical background and theoretical foundation and good generalization performance, Support Vector Machine (SVM) Classification becomes more feasible options for large datasets. A major research goal of SVM is to improve the speed in training and testing phase. In this paper We introduce a proposed algorithm to speed up the training time of SVM is presented. It is highly accurate classification method. However SVM classifiers suffer from slow processing, when training with a large set of data tuples. Our novel approach selects a small representative amount of data from large datasets to enhance training time of SVM. This method uses an induction tree to reduce the training dataset for SVM classification, it generate faster results with improving accuracy rates than the current SVM implementations.

Keywords: Microarray; large datasets; Classification; Decision Tree, SVM

1. Introduction

SVM grew out of early effort by Vladimer Vapnik and Alexei Chervonenkis on statistical learning theory[1] . And the first paper on SVMs was presented by Boser, Guyon, and Vapnik[2]. The SVM is a highly accurate classification method. However SVM classifiers suffer from slow processing, when training with a large set of data tuples. Although the training time of even the fastest SVMs can be extremely slow. Conversely, it is known that the major drawback of SVM occurs in its trainingphase[8, 9]. To investigate this problem the efficient Decision Tree (DT) based

* Corresponding author. Tel.: 9566636669

*E-mail address:*josekiruba@gmail.com

Selection and Peer-review under responsibility of International Conference on Processing of Materials, Minerals and Energy (July 29th – 30th) 2016, Ongole, Andhra Pradesh, India.

Data selection and Support Vector Classification technique is needed for training large data sets, for data reduction DT s have been used in several works .In each disjoint region exposed by a DT to train a SVM[11]. Extracting bioinformatics information from DNA micro array technology is the popular method in biological data. In bioinformatics, SVM classification is one of the popular tool for identify correlation among various clause and the features of various objects. In recent years, determining the informative genes that cause cancer has its great impact in microarray technology. The foremost snag that exists in microarray data analysis is the curse of dimensionality, this issue carried out in many ways. According to their strong mathematical background and generalization capability, SVM produced better result. Meanwhile in its expensive computation, high dependency on the size of input dataset. We proposed a method to reduce the dimensionality of training set for SVM based decision tree. The novel method produces better accuracy and in a faster way that the traditional SVM implementations.

Classification has been an indispensable premise in statistics, data mining, machine learning, bioinformatics and Medical science. Recent advances in data mining research have led to the progress of abundant and scalable methods for mining fascinating patterns and knowledge in large databases. In the past decades spectator changes in biomedical research an explosive progress of biomedical data. In particular cancer therapy investigations find whether it is cancerous cell or not. Our novel approach has training dataset with two classes. SVM is usually regard as the most accurate classification tool for many bioinformatics applications .however the complexity of training an SVM is $O(N^2)$,where N is the number of objects/points. In this progress find how to scale up SVMs for large data novel method uses. Therefore the novel method uses in its training phase with decision tree to reduce the training dataset for SVM. In this paper to reduce the size of datasets based on a data selection method Decision Tree approach proposed. That is ability to deal with redundant attributes and robustness to noise, able to partition the input space into regions with low entropy and liable human interpretation. The experimental results show that proposed method reduces the training time with higher accuracy. Decision Tree (DT) have been used a dimensionality reduction approach in some previous work .It reduces the number of data dimensions and introduce the problems of feature selection and feature construction. In each disjoint region exposed by a decision tree is used to train SVM. Thus the region found by small datasets is less sophisticated than the region obtained by the entire training set. Even though small learning datasets reduce decision tree complexity through decision rule. A SVM is a good classification method was obtainable in[4].It reveal in reducing the number of instances to train a SVM. The key idea was to fairly accurate decision boundary of SVM by using DT that is to confine the objects near the decision boundary. The novel approach presented in this paper is very efficient with biological datasets, first it eliminates dispensable data, and then it recovers the data points that are near to the decision boundaries. The training dataset of SVM is done on those remained valuable data. This effective way to reduce SVM training time approach is necessary a rational tradeoff between accuracy and training time.

2. Related Work

Our novel method is based on that, the region originate on undersized data sets are less elaborated than the region obtained by the full training set[13,14].A related approach to our proposed technique was presented in[16], it consists to reduce the training instances of SVM. The idea was to fairly accurate the decision boundary of SVM by using DT. It also used to choose an example which is near to decision boundary[17], but the consideration of sigma parameters optimization is not proper. The novel method presented in this paper is efficient with large data sets .It eradicates nonessential data, then it recuperate the data points that are near to the decision boundaries. After that the training of SVM is made on those remained valuable data. The way presented in this paper yet prominent and effective mode to reduce the SVM training time. It is very necessary applications on the real world data base accuracy and training time.

3. Proposed Work

The proposed method for classification of microarray input data is processed by statistical analysis and database analysis. To mine microarray and large dataset might require data normalization (data transformation) with esteem to the same control gene and a selection of a subset of treatments (data reduction).In this paper introduce a preprocess step to rapidly remove data that do not contribute to classify the decision boundary of SVM. While

preserving Support vector candidate's .The novel approach applies SVM on a tiny subset of original data set in order to attain a draft of the optimal separating hyperplane, then it labels objects that are remote from sketched hyperplane and objects that are close to it. In this to identify objects that have akin characteristics to the computed Support Vector(SV) and removes not as much of important objects from the original dataset by using a decision tree. In general the datasets contain comparable number of examples with positive and negative labels.SVM can achieve very good accuracy. In some cases, particularly cancer data sets not awfully balanced. There are named as imbalanced or skewed. To overcome this problem, it is necessary steps taken to balance the training set artificially. A mechanism built in the proposed algorithm realizes the balancing task. Due to this progress a small subset C from Entire Data Set (EDS) taking account into the number of elements of each class. The proposed algorithm begins the selection of objects by computing the rate of positive and negative labels. The initial subset selection algorithm mines a small subset C from the entire dataset. Subset C is a parameter that controls tradeoff between margin and error. It computes a outline of the optimal separating hyper plane it maximizing the margin between two classes, the closest samples being indicated as support vectors. The process uses a decision tree to identify objects that have analogous characteristics to the computed SV and then it eradicate the less important objects from the original microarray dataset .The proposed algorithm follows the step procedure as shown in Fig 1.That summarily includes the following

1. Initial subset selection
2. Discover objects close to decision boundaries.
3. Modeling the allocation of support vectors
4. Training the SVM

Initial subset selection extracted a small subset c from the entire data set.

$$EDS = \{(x_1, y_1), i = 1,2 \dots n\} \quad (1)$$

Formally, our training dataset EDS consists of n pairs$(x_1,y_1),(x_2,y_2)\dots(x_n,y_n)$ with $x_i \in R^d$ and $y_i \in \{-1,1\}$ after that apply

$$c = \{x_i, y_i\}_{i=1}^{l} \quad (2)$$

Mainly the function of decision tree is to reduce the entire data set, attain a small data set with the most important data of microarray. here a decision tree in this proposed model to detect and preserve all data are close to the objects $(x_k,y_k) \in V_{smal.}$ To find a decision function that separate SV and Non SV, a decision tree is encouraged. In this method objects close from hyperplane and data points faraway from hyperplane are distinguished. The Reduced Dataset (RDS) as

$$RDS=\{x_{pi,}y_{pi}\} \cup \{\{(x_k,y_k)\} \in V_{small}\}$$

In this approach, the novel algorithm reduces size of large training set. It utilizes a SVM twice. At first, it recovers adequate statistical information from SV and present it to a DT, The entire Data set uses these information and to recover all SV candidates..The second time SVM used to refines the solution to obtain the optimal solution. In order to condense the entire dataset, the following two characteristic be considered.

i. Imbalance of the training set
ii. Size of the training set

In balanced Data sets, the selection progression is random. Besides, the training time rely on the appropriate choice of Parameter σ of the RBF kernel from the grid search. It implements by John Platt's SMO Algorithm for training support vector classifier[20] using weka classifiers, functions support Vector RBF Kernel.

It represent as K(x, y) = e^-(gamma * <x-y, x-y>), Gamma -- The Gamma value 0.01, c 1.0, cache size 250007, Eepsilon value 1.0E-12.

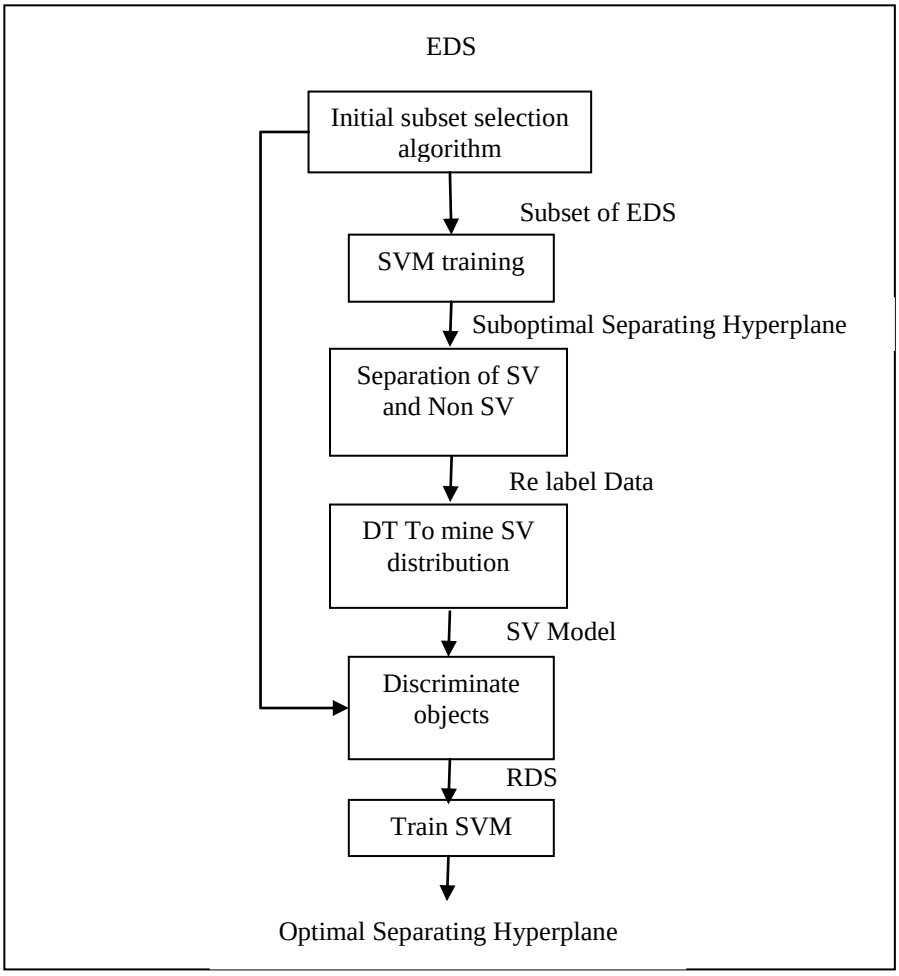

Fig. 1.Block Diagram of the Proposed System

Where x_{pi} are the data close from the decision hyperplane obtained. V_{small} are the support vector from small data set.

$$p \leftarrow \frac{\min(N+,N-)}{|EDS|} \tag{3}$$

$$p_h < \tau_u \quad (4), \quad \tau_u \leq p_m < \tau_b \quad (5), \quad \tau_b \leq p_s \leq 0.5 \quad (6)$$

The thresholds used to indicate when the dataset is considered with small imbalance (6), moderate imbalance (5)and high imbalance (4).and $\tau_b = 0.25$ and $\tau_u = 0.1$

SVM structured as a two class problem, where the classes are separable linearly. The input dataset D be represent in 2D as (x_1, y_1), (x_2, y_2)…. $(x_{|D|}, y_{|D|})$, where x_i is the set of training tuples and y_i is the class label associated with training sample. In a training sample SVM constructs a line of separation for two attributes (x, y) and a plane of separation for three attributes and a hyper plane of separation for n dimensions. To make the SVM optimization problem accurately obedient by writing Minimize in Equation (7), where $\varepsilon_i > 0, i = 1,2 \dots l$

$$\frac{1}{2} w.w + C \sum_{i=l}^{l} \Phi(\varepsilon_i) \tag{7}$$

Algorithm : Initial data selection algorithm
Input data :
The entire Data set EDS, Threshold τ_a, τ_b
Output Data:
EDS_r, Subset (EDS)

Begin IDSA
To compute the value of p using Eqn (3)
If(EDS is balance) **then**
Compute Eqn (6)
Apply Simple random sampling to create reduced EDS_r
Else if EDS is partially balanced then
Compute Eqn (5),and selection of majority and minority classes
Else
The dataset is imbalanced use Eqn (4)
Create EDS_r selection of 100% number of instances of minority, majority classes.
End if
Return EDS_r
End IDSA

Algorithm 1. Initial Data Selection Algorithm (IDSA)

4. Experimental results and Comparison

In order to demonstrate the efficiency of the proposed method, it was tested with micro array cancer data set leukemia, Duke breast cancer, colon, Wisconsin Progonostic Breast Cancer(WPBC),Wisconsin Diagnostic Breast Cancer(WDBC),Gisette,IJCNN1,Epsilon,Susy,KDDCup2010, and Higgs data be considered. The subsequent datasets are used in the experiments.https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html l from UCI machine learning repository. The proposed method was implemented in weka3.7.13. The experiments are agreed on several datasets using Sequential minimal Algorithm in support vector Machine. To train a SVM involves some parameters. It's taking important outcome on the classifier performance. In this paper Radial basis function was used as kernel. In our experiments dataset were normalized and 10 fold cross validation was used to validate the results. The experiments were execute on a 2.1GHz Intel Core i5CPU, 4GB RAM on windows XP.

Table 1.Dataset Description

Data set	Training size	Testing size	Features	Classes
Leukemia	38	34	7129	2
Duke breast cancer	38	4	7129	2
Colon	62	-	2000	2
WPBC	198	-	30	2
WDBC	683	-	10	2
Gisette	6,000	1000	5,000	2
IJCNN1	49,990	91,701	22	2
Epsilon	4,00,000	1,00,000	2,000	2
Susy	50,00,000	5,00,000	18	2
KDDcup2010	84,07,752	5,10,302	202,16,830	2
Higgs	110,00,000	5,00,000	28	2

Table 2.Performance analysis

Data set	SV	Accuracy (Proposed /Traditional SVM)	Time in Sec's (Proposed /Traditional SVM)
Leukemia	17	99.6/98.05	0.21/0.45
Duke breast cancer	15	98.7/96.4	0.19/0.32
Colon	102	96.5/94.5	0.76/0.84
WPBC	65	97.6/98.1	0.65/0.81
WDBC	46	98.1/96.7	0.54/0.65
Gisette	2	99.4./98.3	0.65/0.79
IJCNN1	6,808	99.6/99.4	1.86/9.07
Epsilon	11,243	95.3/92.8	887/1056
Susy	81,423	97.6/96.4	9,581/12,532
KDDcup2010	922,163	98.4/95.5	39,581/64,532
Higgs	119,899	91.45/90.2	80,285/1,78,865

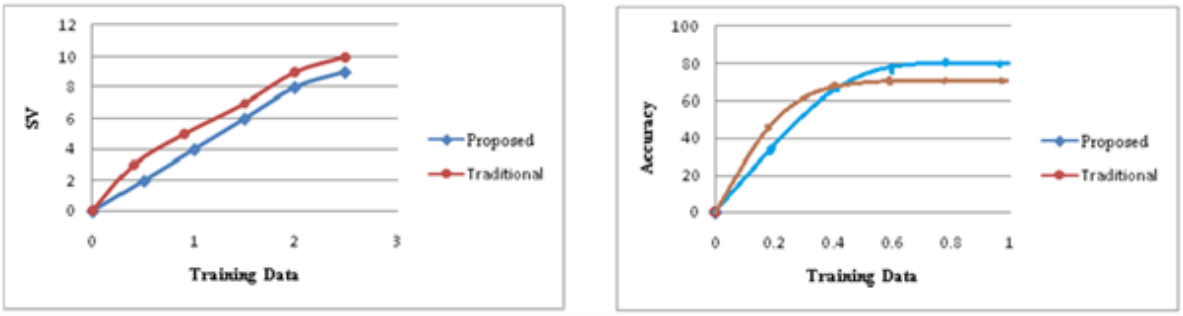

Fig. 2. Performance analysis of micro array data

Fig 2. Shows the performance of the proposed method was compared with traditional SVM. Table 1 shows a microarray dataset features, samples, positive, negative, classes, data size. Table 2 shows the Training size, Testing size, time, accuracy performance of proposed and specified the field with in brackets is the outcome of traditional SVM.

5. Conclusion

In this paper, the preeminence proposed algorithm becomes clearer in micro array gene expression data and large amount of datasets, which reduces size of large training set concept, is proposed. It utilizes a twofold SVM and applies a data filter based on a decision tree, that scans the entire data acquire a small subset of data points. The proposed approach is abstractly simple, easy to implement for our experiments, and faster than other traditional SVM training algorithms. It also captures the pattern of the data and it provides enough information to obtain a good performance. The results of experiments on micro array data set, large data sets show that the proposed approach is scalable for large data classification, while engender high classification accuracy, and effective.

References

[1] Vapnik V N, Chervonenkis A Y,On then Uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications,pp.264-280,(1971).

[2] Boser B, Guyon and Vapnik V N, A training Algorithm for optimal margin classifiers. In Proceeding of fifth Annual Workshop on Computational Learning theory, ACM press, San Mateo.pp.144-152,(1992)

[3] Huang J, Lu J, Ling C, Comparing Naive Bayes, Decision Trees and SVM with AUC and accuracy ,Third IEEE International Conference on Data Mining, ICDM 2003,pp.553-556,(2003).

[4] Lin C F, Wang S D , Fuzzy support vector machines, IEEE Transactions on Neural networks,pp. 464-471, (2002)

[5] Vapnik V N , The nature of statistical learning theory ,Springer-Verlag, (1995).

[6] Burges C J C , A tutorial on support vector machines for pattern recognition, Data mining and Knowledge Discovery,pp.121-167, (1998).

[7] Han-Pang Huang Y H L, Fuzzy Support Vector Machines for Pattern Recognition and Data mining, International Journal of Fuzzy Systems, pp. 826-835, (2002).

[8] Zhang C H, Tian Y J, Deng N Y, The new interpretation of Support Vector Machines on Statistical Learning Theory, Science China Mathematics pp.151-164, (2010).

[9] Deng N Y , Tian Y J, Zhang C H , Support Vector Machines, Optimization Based Theory, Algorithms and Extensions.CRC Press, (2012).

[10] Cristianini N, Shawe-Taylor J, An Introduction to Support Vector Machines and other Kernel based Learning Methods,1st edition. Cambridge University Press,(2000).

[11] Chang F, Guo C Y, Lin X R, Lu C J, Tree decomposition for large scale SVM problems, Journal of Machine Learning,pp.2935-2972,(2010).

[12] Guyon I , Weston J, Barnhill S and Vapnik V, Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning ,pp. 389-422, (2002)

[13] Piramuthu S, Input Data for Decision Trees, Expert System Application,34(2),pp.1220-1226,(2008)

[14] Wang R, He Y L,Chow C Y, Ou F F, Zhang J, Learning ELM tree from big data based on uncertainty reduction, Fuzzy Sets System (2014)

[15] Chih-Chung Chang and Chih-Jen Lin, LIBSVM A library for support vector machines. ACM Transactions on Intelligent Systems and Technology,pp.1–27,(2011). Software available at http://www.csie.ntu. edu.tw/~cjlin/libsvm

[16] Mukherjee S, Tamayo P, Mesirov J P, Slonim D, Verri A, and Poggio T, Support Vector Machine Classification of microarray data. Technical Report 182, (1999).

[17] Arun Kumar M, Gopal M, A hybrid SVM based decision tree, Pattern Recognition,43(12),pp.3977-3987,(2010).

[18] Cervantes J, Lopez A, Garcia F, Trueba A, A fast SVM training Algorithm based on a decision tree data filter,in Advances in Artificial Intelligence ,vol.7094 of Lecture Notes in Computer Science,Springer,pp.187-197,(2011).

[19] Furey S, Nigel Duffy, Nello Cristianini, David Bednarski, Michel Schummer, and David Haussler, Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data. Terrence Bioinformatics. pp.906-914(2000).

[20] Platt J,Fast training of support vector machines using Sequential Minimal Optimization, in Advance kernel methods Support Vector Machine.pp.185-208.(1998)

[21] Chang C. C., and Lin C. J. Training support vector classifiers, Theory and algorithms. Neural Computation vol.13,pp. 214-219,(.2001)

[22] Khan J, et al., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nature Med. vol .7,pp.673 (2001) .

[23] Furey T S, et al., Support Vector Machine Classification and validation of cancer tissue samples using microarray expression data, Bioinformatics, vol .16,pp.906, (2000) .

[24] Pochet N, De Smet F, Suykens J A, De Moor B L, Systematic benchmarking of microarray data classification assessing the role of nonlinearity and dimensionality reduction, Bioinformatics, vol.20,pp.3185 (2004).

[25] Xing E P, Jordan M I, Karp R M, Feature selection for high-dimensional genomic microarray data, in Proceedings of the 18th International Conference on Machine Learning, (2001)

[26] Venkatesh and Thangaraj, Investigation of Micro Array Gene Expression Using Linear Vector Quantization for Cancer", International Journal on Computer Science and Engineering, Vol. 02, No. 06, pp. 2114-2116, (2010).

[27] Ye J, Li T, Xiong T, and Janardan R, Using uncorrelated discriminant analysis for tissue classification with gene expression data, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 1, no. 4, pp. 181–190, (2004).

[28] Peng Y, Li W, and Liu Y, A hybrid approach for biomarker discovery from microarray gene expression data for cancer classification, Cancer Informatics, vol. 2, pp. 301–311, (2007).

[29] Bharathi A and Natarajan A, Cancer classification of bioinformatics data using ANNOVA, International Journal of Computer Theory and Engineering, vol. 2, no. 3, pp. 369–373, (2010).

[30] Peng Y, A novel ensemble machine learning for robust microarray data classification, Computers in Biology and Medicine, vol. 36, no. 6, pp. 553–573, (2006).

[31] Lee C and Leu Y, A novel hybrid feature selection method for microarray data analysis, Applied Soft Computing Journal, vol. 11, no. 1, pp. 208–213, (2011).

[32] Arun Kumar M ,Goal M ,A hybrid SVM based decision tree pattern Recognition.pp.3977-3987(2010).

Selection and Peer-review under responsibility of International Conference on Processing of Materials, Minerals and Energy (July 29th – 30th) 2016, Ongole, Andhra Pradesh, India.

Keywords: Type your keywords here, separated by semicolons ;