Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 7, Issue. 2, February 2018, pg.42 – 48

A HYBRID APPROACH FOR AN EFFICIENT CLASSIFICATION USING DECISION TREE AND SVM

A.K. Shafreen Banu[1], Dr. S. Hari Ganesh[2]

[1]*Asst. Professor, Dept. of Information Technology, Bishop Heber College, Trichy, Tamilnadu, India*
Email: shafreenbanu@gmail.com

[2]*Asst. Professor, Dept. of Computer Science, H.H. The Rajah's College, Pudukottai, Tamilnadu, India*
Email: hariganesh17@gmail.com

Abstract: ***Nowadays real world data bases observed significant growth in the volume of data in digital format, due to the extensive use of datasets and storage system. It is essential for developing fast and accurate algorithms to automatically classify large data. However the data size increases the proposed method make faster computation and scalable machine learning algorithm is used to learn faster from the labeled training data. For a large datasets the Support Vector Machine (SVM) Classification becomes more feasible options. A major research goal of SVM is to improve the speed in training and testing phase. This paper proposed an algorithm to speed up the training time of SVM. SVM is a highly accurate classification method. When training with a large datasets the SVM classifiers suffer from slow processing. The enhanced approach selects a small amount of data from large datasets to enhance training time of SVM. The proposed 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. In this paper, a hybrid approach of classification is proposed which attempts to utilize the advantages of both decision trees and SVM leading to better classification.***

Keywords: Support Vector Machine, Classification, Decision tree, Hybrid approach

I. INTRODUCTION

The Support Vector Machine is a commonly used method for classification and has been used in variety of applications [8]. It is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm each data item can plotted as a point in n-dimensional space with the value of each feature. Support Vectors are simply the co-ordinates of individual observation. The fundamentals of Support Vector Machines based on statistical learning theory used to solve the classification problem. Support Vector Machine is a frontier which best segregates the two

classes (hyper-plane/ line).The SVM is the recent addition to the toolbox of data mining practitioners and are gaining popularity due to many attractive features, and promising experiential performance[6,9].

Decision tree is the efficient tool for classification and prediction [1]. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. The recursive partitioning is used to build the decision tree. A tree can be "learned" by splitting the source set into subsets based on an attribute value test [3]. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursive process is completed when splitting no longer adds value to the predictions or when the subset at a node all has the same value of the target variable. The construction of decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data and also had good accuracy. It can classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance [4]. The decision tree classifies the data items by starting at the root node of the tree, testing the attribute specified by this node and then moving down the tree branch corresponding to the value of the attribute. This process is then repeated for the subtree rooted at the new node.

Integrating decision-making tree and SVM models gives better performance than the individual learning or decision-making models. Even though the SVM is a highly accurate classification method it suffers from slow processing and the training time is extremely slow when training with a large set. The major drawback of SVM occurs in its training phase [8, 9]. To solve this problem the efficient Decision Tree based Data selection and Support Vector Classification technique is needed for training large data sets.

Classification has been an essential basis in statistics, machine learning, data mining, bioinformatics and Medical science [2, 5]. The proposed approach has training dataset with two classes. SVM is usually regard as the most accurate classification tool for many bioinformatics applications [10] .However the complexity of training an SVM is $O(N^2)$,where N is the number of objects [6]. The proposed 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. The experimental results show that proposed method reduces the training time with higher accuracy. Decision Tree reduces the number of data dimensions and introduces the problems of feature selection and feature construction. The each disjoint region exposed by decision tree is used to train the 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.

II. PROPOSED WORK

The complexity of training phase and required storage memory for saving these data will increase accordingly when dataset is large. Therefore, it's necessary to have a model that is able to reduce the complexity. The proposed model is a combination of SVM and C4.5 in order to achieve an efficient integrated procedure for classification. The figure 1 shows the flow chart of the proposed model.

The proposed method involves dividing all data into two groups of experimental and test data in a random way following a proportion of 70 to 30. Then the experimental data are fed into standard SVM, and output is estimated. Data are classified again with the aid of SVM using the obtained coefficients. The estimated class was called new target. As the next step, distance between individual data is obtained by support vectors corresponding to the estimated class then their average value is calculated.

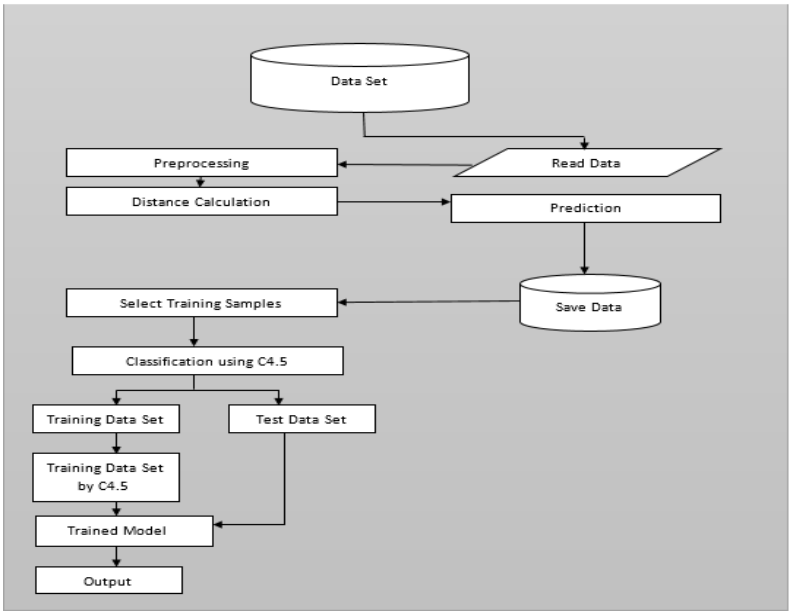

Figure 1: Block Diagram of the Proposed System

The proposed model for data classification includes following steps:

Step 1: Data reading
Step 2: Preprocessing of data (including normalization and removing deviated data)
Step 3: Applying support vector machine on dataset
Step 4: Calculating the distance of sample data for support vectors relevant to each class
Step 5: Predicted label and obtained distance as data membership in one class is calculated together with actual label of each sample class.
Step 6: The obtained results from step 5 are stored into a new dataset.
Step 7: Classification of training data is carried out using the new dataset by decision tree.
Step 8: Data testing
Step 9: Evaluating outcomes of the proposed model

The estimated class and the value of calculated distance for each experimental data as the feature vector together with actual class of data are fed into the decision tree classifier to get the results recalculated. The above steps are repeated during test phase in such a way that individual test data are initially obtained with the aid of SVM model. Then they are classified, and the estimated

class for each test data is considered as the new target. As the next step, distance between test data from support vectors corresponding to new target is averaged, and the resultant value is considered as the second feature. Two obtained features (estimated class and distance) are fed into the already acquired decision tree classifier in order to verify the class of test data.

III. RESULTS AND ANALYSIS

A. Data Description

To evaluate the SVM-DT model 5 datasets from the UCI repository and 1 from Statsoft STATISTICA dataset examples. The datasets used in this study are Zoo, Wine, Pima Indian, Iris, Ionosphere and Leukemia. The table 1 describes the characteristics of the datasets.

Table 1: Description of Datasets

Dataset	No. of Attributes	No. of Instances
Zoo	18	101
Wine	14	178
Pima Indian	15	690
Iris	4	150
Ionosphere	34	351
Leukemia	4	71

B. Experimental Setup

All the datasets have been classified using three classifiers namely C4.5, SVM and the proposed model to study the performance of the new proposed model. 10x10 cross-validations have been employed to estimate the classification accuracies of the tree models. The experiments have been carried out on WEKA. C4.5 tree has been built in WEKA and for SVM classification, STATISTICA has been used.

C. Results

The experiments have been carried out and the results obtained have been encouraging. The results for the proposed model are presented in Table II. The table II shows the classification accuracy using C4.5, SVM and proposed model.

Table 2: Classification Accuracy Using C4.5, SVM and Proposed Model

Dataset	C4.5	SVM	Proposed Model
Zoo	92.07	95.05	98.18
Wine	93.52	95.43	98.31
Pima Indian	73.82	75.99	79.34
Iris	94.00	95.66	97.08
Ionosphere	91.45	92.87	94.14
Leukemia	87.32	90.14	96.96

Table 2 shows the classification accuracies of the proposed model SVM-DT has performed quite well compared to C4.5 and SVM. For decision tree performance evaluation, the number of leaves and the depth of the tree are very important factors as they contribute to the better comprehensibility of the decision tree obtained. From table 3, it can be observed that the number of leaves and the depth of the tree decreases remarkably for the proposed model. For three datasets i.e., zoo, Pima Indian and Leukemia, the proposed model has a tree size one, i.e. the tree doesn't grow beyond the root node. For these datasets, there is only one SVM which classifies it efficiently. Wine and Iris datasets have 7 leaf nodes that means there are 7 SVMs acting as the leaf nodes, whereas Ionosphere has 9 SVMs. This establishes the efficiency of the proposed model with respect to classification accuracy, comprehensibility and time. The proposed model has been applied on comparatively smaller datasets; the behavior of the proposed model may differ when applied on larger datasets. It is observed that the proposed model has yielded higher accuracy for all the datasets as compared to C4.5. The table 4, shows that the time taken by the C4.5, SVM and the proposed model.

Table 3: Number of Leaves and Tree Size Using C4.5 and Proposed Model

Dataset	C4.5		Proposed Model	
	No. of Leaf Node	Tree Size	No. of Leaf Node	Tree Size
Zoo	9	17	5	10
Wine	5	9	4	7
Pima Indian	20	39	6	8
Iris	5	9	4	7
Ionosphere	18	35	5	9
Leukemia	4	7	2	2

Table 4: Time Accuracy Using C4.5, SVM and Proposed Model

Data Set	Time in Sec's		
	C4.5	SVM	Proposed Model
Zoo	0.45	0.31	0.25
Wine	0.39	0.35	0.15
Pima Indian	0.76	0.68	0.53
Iris	0.54	0.44	0.35
Ionosphere	0.65	0.63	0.53
Leukemia	0.59	0.45	0.32

Figure 2 depicts a diagram comparing classification accuracy of C4.5, SVM and proposed models. Figure 3 depicts a diagram comparing time taken by C4.5, SVM and proposed models.

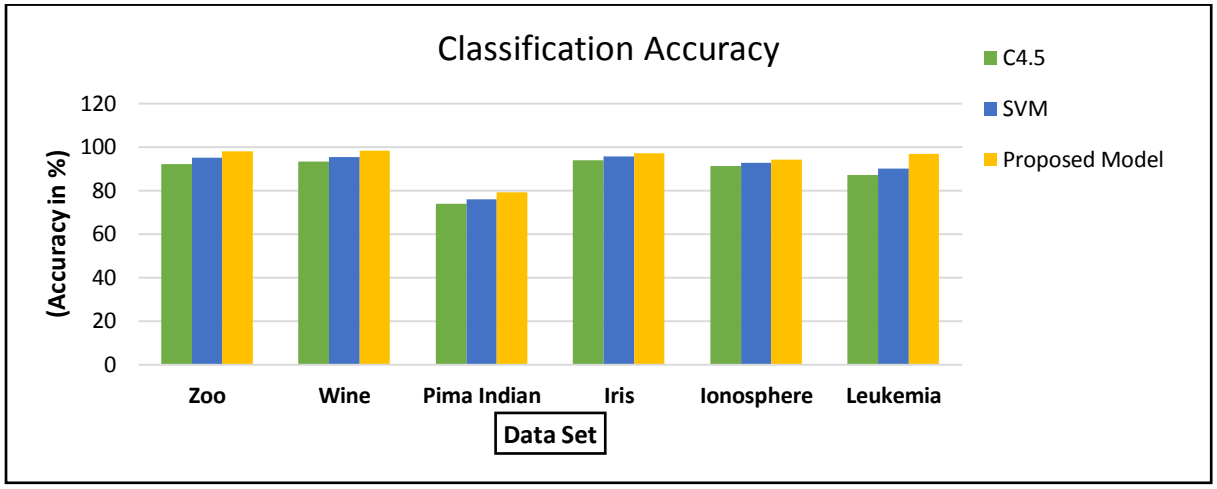

Figure 2: Classification Accuracy of C4.5, SVM and Proposed Model

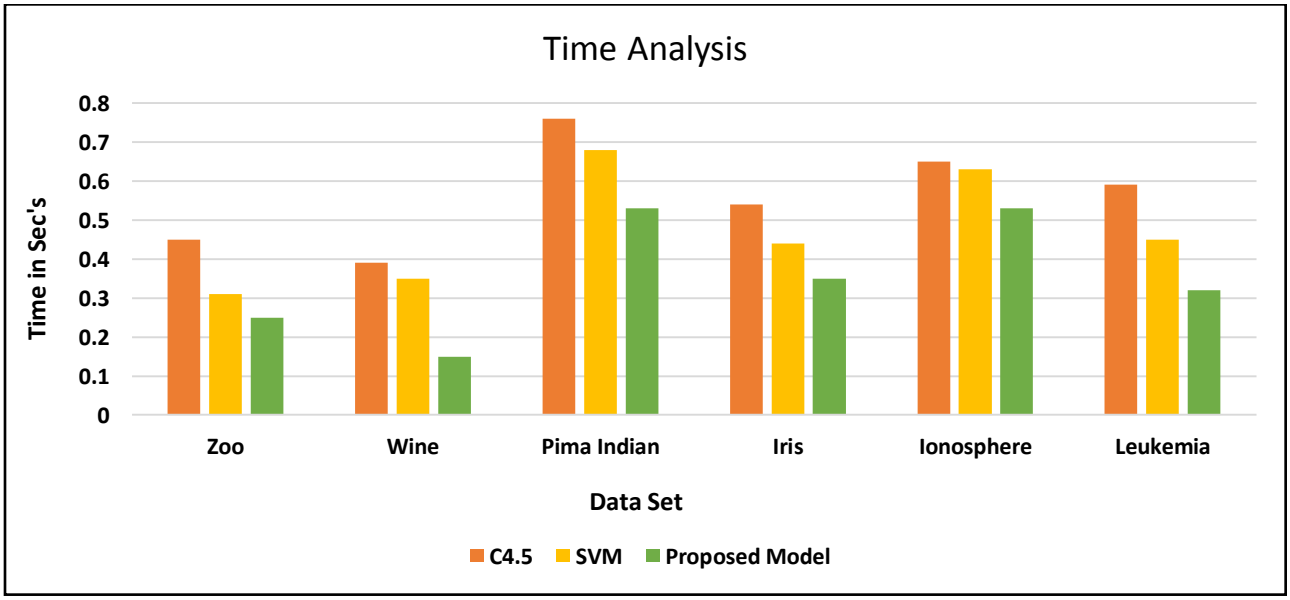

Figure 3: Time Analysis of C4.5, SVM and Proposed Models

IV. CONCLUSION

In this paper, the proposed algorithm becomes clearer and reduces size of large training set. The proposed approach is conceptually 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 show that the proposed approach is scalable for large data classification, while engender high classification accuracy, and effective.

REFERENCES

[1] Badr Hssina, Abdelkarim Merbouha, Hanane Ezzikouri, Mohammed Erritali, "A comparative study of decision tree ID3 and C4.5", (IJACSA).

[2] Devinder Kaur, Rajiv Bedi and Dr. Sunil Kumar Gupta, "Implementation of Enhanced Decision Tree Algorithm on Traffic Accident Analysis", (IJSRT), ISSN: 2379-3686, 15th September 2015.

[3] Sonali Agarwal, G. N. Pandey, and M. D. Tiwari, "Data Mining in Education: Data Classification and Decision Tree Approach", International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 2, No. 2, April 2012.

[4] Heling Jiang, An Yang , Fengyun Yan and Hong Miao, "Research on Pattern Analysis and Data Classification Methodology for Data Mining and Knowledge Discovery" , International Journal of Hybrid Information Technology Vol.9, No.3 (2016), pp. 179-188

[5] V. Karpagam, and R. Rangarajan, "Improved content-based classification and retrieval of images using support vector machine", CURRENT SCIENCE, VOL. 105, NO. 9, 10 NOVEMBER 2013.

[6] Begum S, Chakraborty D, and Sarkar R., " Data classification using feature selection and KNN machine learning approach", In the International Conference on Computational Intelligence and Communication Networks, IEEE, Jabalpur, India: 811-814, (2015).

[7] Vijayan A, Kareem S, and Kizhakkethottam JJ., "Face recognition across gender transformation using SVM classifier", Procedia Technology, 24: 1366-1373, (2016).

[8] Ahmed, Karim , Mohamed, and Nacéra , "A Hybrid Approach for Automatic Classification of Brain MRI Using Genetic Algorithm and Support Vector Machine", Leonardo Journal of Sciences ISSN 1583-0233 Issue 17, July-December 2010.

[9] Aparna P K , Dr. Rajashree Shettar, "Hybrid Decision Tree using K-Means for Classifying Continuous Data" , International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 10, ISSN(Online): 2320-9801,October 2015.

[10] Abhishek Bargaje, Shubham Lagad, Ameya Kulkarni, and Aniruddha Gokhale, "Review of Classification algorithms for Brain MRI images" , International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 01 , -ISSN: 2395 -0056Jan-2017

