Hybrid Approach for Classification using Support Vector Machine and Decision Tree

Anshu Bharadwaj
Indian Agricultural Statistics research Institute
New Delhi, India
anshu@iasri.res.in

Sonajharia Minz
Jawaharlal Nehru University
New Delhi, India
minz@jnu.ac.in

Abstract— **A hybrid system or hybrid intelligent system uses the approach of integrating different learning or decision-making models. Each learning model works in a different manner and exploits different set of features. Integrating different learning models gives better performance than the individual learning or decision-making models by reducing their individual limitations and exploiting their different mechanisms. 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 accuracy.**

Keywords—*hybrid, support vector machine, decision tree, ID3, C4.5*

I. INTRODUCTION

Integrating different learning models gives better performance than the individual learning or decision-making models by reducing their individual limitations and exploiting their different mechanisms. In a hierarchical hybrid intelligent system each layer provides some new information to the higher level [1]. The overall functioning of the system depends on the correct functionality of all the layers. A hybrid system or hybrid intelligent system uses the approach of integrating different learning or decision-making models. Each learning model works in a different manner and exploits different set of features. Given a classification problem, no one classification technique always yield the best results, therefore there have been some proposals that look at combining techniques.

i. A synthesis of approaches takes multiple techniques and blends them into a new approach.
ii. Multiple independent approaches can be applied to a classification problem, each yielding its own class prediction. The results of these individual techniques can then be combined in some manner. This approach has been referred to as combination of multiple classifiers (CMS).
iii. One approach to combine independent classifiers assumes that there are n independent classifiers and that each generates the posterior probability.

Support vector machine is a widely used method for classification and have been used in variety of applications. The foundations of Support Vector Machines (SVMs) based on statistical learning theory have been developed by [21], [6] to solve the classification problem. The support vector machine (SVM) is the recent addition to the toolbox of data mining practitioners and are gaining popularity due to many attractive features, and promising empirical performance. They are a new generation learning system based on the latest advances in statistical learning theory. The formulation embodies the Structural Risk Minimization (SRM) principle, which has been shown to be superior [19], to traditional Empirical Risk Minimization (ERM) principle. Decision Tree [7], [10] is commonly built by recursive partitioning. A univariate (single attribute) split is chosen for the root of the tree using some criterion (e.g., mutual information, gain ration, gini index). The data is then divided according to the test, and the process repeats recursively for each child. After a full tree is built, a pruning step is executed, which reduces the tree size.

In this paper, a hybrid approach is proposed using decision tree and support vector machine. The hybrid model proposed attempts to embed SVM within a C4.5 algorithm of decision tree as a decision tree pre-pruning method and resulting into a more accurate and efficient hybrid classifier.

II. PRELIMINARIES

A. Decision Tree Algorithm: C4.5

C4.5 belongs to a succession of decision tree learners that trace their origins back to the work of Hunt and others in the late 1950s and early 1960s [3]. C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan [7]. C4.5 is an extension of Quinlan's earlier ID3 algorithm. C4.5 made a number of improvements to ID3. Some of these are:

i. Handling both continuous and discrete attributes - In order to handle continuous attributes, C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it [8].

ii. Handling training data with missing attribute values - C4.5 allows attribute values to be marked as ? for missing. Missing attribute values are simply not used in gain and entropy calculations.
iii. Handling attributes with differing costs.
iv. Post-Pruning - C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.

ID3 approach favours the attributes with many divisions and thus may lead to over-fittng. An improvement can be made by taking into account the cardinality of each division. This approach uses the GainRatio as opposed to Gain. For splitting purpose, C4.5 uses the largest GainRatio that ensures a larger than average information gain. This is to compensate for the fact that GainRatio value is skewed toward splits where the size of one subset is close to that of the starting one.

B. Support Vector Machine

SVM belongs to the class of supervised learning algorithms in which the learning machine is given a set of examples (or inputs) with the associated labels (or output values). Like in decision trees, the examples are in the form of attribute vectors, so that the input space is a subset of Rn. SVM is a classifier that searches for a hyperplane with the largest margin, which is why it is known as maximum margin classifier. SVMs create a hyperplane that separates two classes (this can be extended to multi class problems). While doing so, SVM algorithm tries to achieve maximum separation between the classes. Separating the classes with a large margin minimizes a bound on the expected generalization error. By "minimum generalization error", it means that when new examples (data points with unknown class values) arrive for classification, the chance of making error in the prediction (of the class to which it belongs) based on the learned classifier (hyperplane) should be minimum. Intuitively, such a classifier is one which achieves maximum separation-margin between the classes. The two planes parallel to the plane are called bounding planes. The distance between these bounding planes is called margin and by SVM "learning", i.e. finding hyperplane which maximizes this margin. The points (in the dataset) falling on the bounding planes are called the support vectors. SVM has greater advantages over other classifiers since they are independent of the dimensionality of the feature space. Use of quadratic programming in SVM has an edge over other classifiers which gives only local minima whereas SVM provides global minima. But at the same time SVM also has a limitation of not considering spatial autocorrelation while classifying the data. SVM was designed initially as binary classifier i.e. it classifies the data into two classes but researchers have extended its boundaries to be a multi-class classifier. SVM was first introduced as a training algorithm that automatically tunes the capacity of the classification function maximizing the margin between the training patterns and the decision boundary [14]. This algorithm operates with large class of decision functions that are linear in their parameters but not restricted to linear dependences in the input components. For the computational considerations, SVM works well on the two important practical considerations of classification algorithms i.e. speed and convergence.

C. Decision Tree Pre-pruning

Decision trees generated by methods such as ID3 and C4.5 are considered to be accurate and efficient, they often suffer the disadvantage of providing very large trees that make them incomprehensible to experts [9]. Tree pruning methods address this problem as well as the problem of over-fitting the data. Such methods typically use statistical measures to remove the least reliable branches. Pruned trees tend to be smaller and less complex and, thus, easier to comprehend. Pruned trees tend to be smaller and less complex and, thus, easier to comprehend [4]. Tree pruning are methods have two approaches: pre-pruning and post-pruning. In the pre-pruning approach, a tree is "pruned" by halting its construction early (e.g., by deciding not to further split or partition the subset of training tuples at a given node), Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset tuples or the probability distribution of those tuples, whereas post pruning removes subtrees from a "fully grown" tree. A subtree at a given node is pruned by removing its branches and replacing it with a leaf. The leaf is labeled with the most frequent class among the subtree being replaced.

D. Statistical Evaluation of Classifiers

Analysis of differences between the algorithms has always been of great interest. There is a fundamental difference between the tests used to assess the difference between two classifiers on a single data set, differences over multiple data sets and differences between multiple classifiers on multiple datasets. Statistics offers more powerful specialized procedures for testing the significance of differences between multiple classifiers. Two well-known methods are Analysis of Variance (ANOVA) and its non-parametric counterpart, the Friedman test. In this study the non-parametric Friedman test has been used to evaluate the difference between the three classifiers and Wilcoxon Signed Ranks test for two classifiers.

1) Friedman Test: The Friedman test [12], [13] is a non-parametric equivalent of the repeated-measures Analysis of Variance (ANOVA). It ranks the algorithms for each data set separately, the best performing algorithm getting the rank of 1, the second best rank 2. . . , In case of ties, average ranks are assigned.

2) Wilcoxon Signed Ranks Test**:** The Wilcoxon signed-ranks test is named for Frank Wilcoxon (1892–1965) [5] is a non-parametric alternative to the paired t-test, which ranks the differences in performances of two classifiers for each data set, ignoring the signs, and compares the ranks for the positive and the negative differences. The test was popularized by [18].

III. PROPOSED MODEL

The proposed method is a hybrid approach to embedding SVM in Decision Tree (SVM-DT) for pre-pruning the tree while carrying out the classification. This resulting hybrid system is categorized as embedded hybrid system where the technologies participating are integrated in such a manner that they appear to be inter-twined. The proposed model is similar to the classical recursive partitioning schemes, except that the leaf nodes created are Support Vector Machine categorizers instead of nodes predicting a single class. The SVM classifier has been used for pre-pruning the DT resulting in a smaller DT than a complete on application of C4.5.

The proposed model uses the C4.5 algorithm for constructing a decision tree. Root node of the decision tree is selected based on a chosen threshold value of the continuous attribute. For this the standard entropy minimization technique is used. In the next step the Significance of Node is computed by using 10x10 cross-validation accuracy estimates for SVM at the node. Computation of Significance of Node is followed by the computation of Significance of Split. The Significance of Split is computed by taking the weighted sum of the significance of the nodes. Here, the weight given to a node is proportional to the number of instances that go down to that node.Significance of Node and Significance of Split are computed and compared and the results attempt to approximate whether the generalization accuracy for SVM classifier at each leaf is higher than a single SVM classifier at the current node. A split is defined to significant if the relative (not absolute) reduction in error is greater than 5% and there are at least 20 instances in the node. If there are n training samples, and m attributes, then the computational complexity of the algorithm for the proposed model has been worked out to be *O(m.n2).*

The resulting model resembles the Utgoff's Perceptron trees [15], the difference is in the induction process. Kohavi [17], proposed an algorithm, which induces a hybrid of decision tree classifiers and Naïve Bayes classifiers.

IV. EXPERIMENTS USING PROPOSED MODEL

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:

1. Zoo
2. Wine
3. Pima Indian
4. Iris
5. Ionosphere
6. Leukemia

To explore the applicability SVM as a tree pruning technique, the datasets used in this study have been comparatively smaller in size with respect to number of instances, i.e., the number of instances are less and not very large. The largest dataset has 690 instances and the smallest has 71. Table I describes the characteristics of the datasets.

TABLE I. 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-validation has been employed to estimate the classification accuracies of the tree models. The experiments have been carried out on WEKA 3.6.5 and STATISTICA Data Miner. C4.5 tree has been built in WEKA and for SVM classification, STATISTICA has been used. SVM-DT has been applied by using the combination of the two softwares. For SVM, RBF kernel has been used and the model parameters have been selected using the grid search method. The evaluation of the classifiers has been done using statistical tests: the Friedman test for multiple classifiers and Wilcoxon Signed Rank test for two classifiers.

V. RESULTS, EVALUATION AND ANALYSIS

A. 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.

TABLE II. CLASSIFICATION ACCURACY USING C4.5 AND SVM-DT

Dataset	C4.5	SVM	SVM-DT
Zoo	92.07	95.05	98.18
Wine	93.82	98.43	98.31
Pima Indian	73.82	77.99	75.34
Iris	96	98.66	97. 08
Ionosphere	91.45	92.87	94.14
Leukemia	87.32	90.14	88.96

As exhibited in table II, the classification accuracies show that the proposed model SVM-DT has performed quite well. The accuracies of all the datasets, except for Pima Indian, have

gone up for SVM and proposed model SVM –DT as compared to C4.5. Whereas, SVM performs very efficiently in terms of classification accuracy as compared to C4.5, proposed model SVM-DT still outperforms the two classifiers except for Pima Indian where the proposed model has not performed better than SVM and C4.5.

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 III, 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 effieciently. 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 SVM-DT has yielded higher accuracy for all the datasets as compared to C4.5 but SVM has shown better performance for Wine, Pima Indian, Iris and Leukemia datasets For Zoo and Ionoshpere, SVM-DT has outperformed SVM. Even if SVM-DT has not performed better than SVM for some datasets, still the depth of the tree has reduced considerably resulting in less time taken and better comprehensibility.

TABLE III. NO. OF LEAVES AND TREE SIZE USING C4.5 AND SVM-DT

Dataset	C4.5		Proposed Model	
	No. of Leaf node	Tree Size	No. of Leaf node	Tree Size
Zoo	9	17	1	1
Wine	5	9	4	7
Pima Indian	20	39	1	1
Iris	5	9	4	7
Ionosphere	18	35	5	9
Leukemia	4	7	1	1

B. Evaluation

1) *Friedman test for multiple classifiers:* For carrying out the statistical analysis of the three classifiers, C4.5, SVM and SVM-DT, Friedman test has been used. The accuracy of the three classifiers has been used to calculate the Friedman test statistic and then obtained value of Friedman test statistic is compared with the critical value Friedman test statistic at 0.05 level of significance. The obtained value of Friedman test statistic for the three classifiers i.e. the proposed model and SVM is -6.00, and the critical value of Wilcoxon test statistic at N=6, k=3 and =0.05 is 7.00, since the obtained value is quite less than the critical value, it has been concluded that the difference between the three classifiers are significantly different i.e. the difference between their performance is unlikely to occur by chance.

2) *Wilcoxon Signed-Ranks test for two classifiers:* The two set of two classifiers each i.e. (C4.5, SVM-DT) and (SVM, SVM-DT), have been evaluated using the Wilcoxon Signed-Ranks test. The accuracy of both the classifiers in both sets has been used to calculate the Wilcoxon test statistic and then obtained value of Wilcoxon test statistic is compared with the critical value Wilcoxon test statistic at 0.05 level of significance. The obtained value of Wilcoxon test statistic for the C4.5 and SVM-DT classifiers i.e. the proposed model and C4.5 is -4.4028, and the critical value of Wilcoxon test statistic at N=6 and =0.05 is 0.00, since the obtained value is quite less than the critical value, it has been concluded that these two classifiers are significantly different i.e. the difference between their performance is unlikely to occur by chance. Similarly, the obtained value of Wilcoxon test statistic for the SVM and SVM-DT classifiers i.e. the proposed model and SVM is -7.23317, and the critical value of Wilcoxon test statistic at N=6 and =0.05 is 0.00, since the obtained value here for these two is also quite less than the critical value, it has been concluded that the difference these two classifiers are also significantly different i.e. the difference between their performance is unlikely to occur by chance.

VI. CONCLUSION

Decision trees are non-parametric estimators and can approximate any "reasonable" function as the database size grows [11]. In practice, it is seen that some parametric estimators such as SVM, may perform better. SVMs can learn a larger set of patterns and be able to scale better, because the classification complexity does not depend on the dimensionality of the feature space. SVMs also have the ability to update the training patterns dynamically whenever there is a new pattern during classification.

The resulting classifier is as easy to interpret as decision-trees and Support Vector Machines. The decision-tree segments the data, a task that is considered to be an essential part of the data mining process in large databases [16]. Each segment of the data, represented by a leaf, is described through a Support Vector Machines.

REFERENCES

[1] A. Abraham, K. Mario "Hybrid Information Systems, First International Workshop on Hybrid Intelligent Systems, Adelaide, Australia, Proceedings Physica-Verlag. 2000

[2] B.E.Boser, I.M.Guyon, and V.N. Vapnik. "A Training Algorithm for Optimal Margin Classifiers". Proceedings of 5th Annual Workshop on Computer Learning Theory, Pittsburgh, PA: ACM, pp.144-152.1992.

[3] E.B. Hunt, Marin. and P.J.Stone. Experiments in induction, Academic Press, New York. 1966.

[4] J.Han and M. Kamber. Data Mining Concepts and Techniques, Morgan Kaufman Publishers.2006.

[5] F. Wilcoxon. Individual comparisons by ranking methods, Biometrics Bulletin, 1, 80-83.1945.

[6] J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery. Vol. 2, pp121-167, 1998

[7] J.R.Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.1993.

[8] J.R.Quinlan. "Improved use of continuous attributes in c4.5". Journal of Artificial Intelligence Research, 4:77-90, 1996.

[9] J.R.Quinlan. "Simplifying decision trees", Int. J. Human-Computer Studies, (1999)51, pp. 497-491, 1999.

[10] L. Breiman, J.H.Friedman, R.A. Olshen and C.J..Stone "Classification and Regression Trees", Wadsworth International Group, Belmont, California, 358 pp. 1984.

[11] L. Gordon and R.A. Olshen. "Almost sure consistent nonparamateric regression from recursive partitioning schemes". Journal of multivariate analysis. 15. pp. 147-163. 1984.

[12] M.Friedman. "A comparison of alternative tests of significance for the problem of m rankings". Annals of Mathematical Statistics, 11: pp. 86–92. 1940.

[13] M.Friedman. "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance", Journal of the American Statistical Association, 32, pp. 675-701.1937.

[14] N.Cristianini and J. Shawe-Taylor. "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", Cambridge University Press, 2000.

[15] P.E.Utgoff. Perceptron Trees: a case study in hybrid concept representation. In Proceedings of the seventh national Conference on Artificial Intelligence, 601-606. 1998.Morgan Kaufmann.

[16] R. Brachman, T. Khabaza, W.Kloesgan, G.Piatetsky-Shapiro and E. Simoudis, "Mining Business Databases", Comm. ACM, Vol. 39, No. 11, pp. 42-48, 1996.

[17] R. Kohavi. Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid. KDD-96.1996.

[18] S. Siegel. Nonparametric Statistics for the Behavioral Sciences, New York, McGraw-Hill.1956.

[19] S.R.Gunn, M. Brown, and K.M. Bossely. "Network Performance Assessment for Neurofuzzy Data Modelling". Intelligent Data Analysis, Vol. 1208, Lecture Notes in Computer Science (X. Liu, P. Cohen, and M. Berthold (Ed.)), 1997. pp. 313-323.

[20] V. Vapnik and C.Cortes. Support-vector networks, Machine Learning, 20(3) pp.273-297.1995.

[21 V. Vapnik. "Statistical Learning Theory". Wiley, NY, 1998.