iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** XII **Month of publication:** December 2014

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Performance Analysis of Automobile Data using Bagged Ensemble Classifiers

M.Govindarajan[#1], A.Mishra[*2]
[#1] *Assistant Professor, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar – 608002, Tamil Nadu*
[*2] *Professor, Department of Mechanical Engineering, Indira Gandhi Institute of* Technology, *Sarang, Odisha*

Abstract— Data mining is the use of algorithms to extract the information and patterns derived by the knowledge discovery in databases process. Classification maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data. The feasibility and the benefits of the proposed approaches are demonstrated by the means of Auto imports and Car Evaluation Databases. A variety of techniques have been employed for analysis ranging from traditional statistical methods to data mining approaches. Bagging and boosting are two relatively new but popular methods for producing ensembles. In this work, bagging is evaluated on Auto Imports and Car Evaluation Databases in conjunction with radial basis function and support vector machine as the base learners. The proposed bagged radial basis function and support vector machine is superior to individual approaches for Auto imports and Car Evaluation Databases in terms of classification accuracy.
Keywords— Data Mining, Support Vector Machine, Radial Basis Function, Classification Accuracy, Ensemble Method

I. INTRODUCTION

Data mining methods may be distinguished by either supervised or unsupervised learning methods. In supervised methods, there is a particular pre-specified target variable, and they require a training data set, which is a set of past examples in which the values of the target variable are provided. Classification is a very common data mining task. In the process of handling classification tasks, an important issue usually encountered is determining the best performing method for a specific problem. Hybrid models have been suggested to overcome the defects of using a single supervised learning method, such as radial basis function and support vector machine techniques. Hybrid models combine different methods to improve classification accuracy. The combined model is usually used to refer to a concept similar to a hybrid model. Combined models apply the same algorithm repeatedly through partitioning and weighting of a training data set. Combined models also have been called Ensembles. Ensemble improves classification performance by the combined use of two effects: reduction of errors due to bias and variance.
The goal of ensemble learning methods is to construct a collection (an ensemble) of individual classifiers that are diverse and yet accurate. If this can be achieved, then highly accurate classification decisions can be obtained by voting the decisions of the individual classifiers in the ensemble.
Two of the most popular techniques for constructing ensembles are bootstrap aggregation [1] and the Adaboost family of algorithms [7]. Both of these methods operate by taking a base learning algorithm and invoking it many times with different training sets.
The rest of this paper is organized as follows: Section 2 describes the related work. Section 3 presents hybrid intelligent system and Section 4 explains the performance evaluation measures. Section 5 focuses on the experimental results and discussion. Finally, results are summarized and concluded in section 6.

II. RELATED WORK

Data mining tasks like clustering, association rule mining, sequence pattern mining, and classification are used in many applications. Some of the widely used data mining algorithms in classification include Support vector machines and neural networks.
Support vector machines (SVMs) are relatively new techniques that have rapidly gained popularity because of the excellent results they have achieved in a wide variety of machine learning problems, and because they have solid theoretical underpinnings in statistical learning theory [5].
On the other hand, Artificial Neural Networks (ANN) as a classifier algorithm are also widely-used in data mining for performing classification in a number of applications. Reference [6] uses ANN and compares its performance against decision trees mining algorithm to develop a prediction models for breast cancer. Reference [12] performs a comparison between ANN and Support Vector Machine (SVM) for Drug/Nondrug Classification.

The ensemble technique, which combines the outputs of several base classification models to form an integrated output, has become an effective classification method for many domains ([8], [10]).
Reference [2] showed that bagging is effective on "unstable" learning algorithms where small changes in the training set result in large changes in predictions. Reference [2] claimed that neural networks and decision trees are example of unstable learning algorithms.
The boosting literature [14] has recently suggested (based on a few data sets with decision trees) that it is possible to further reduce the test-set error even after ten members have been added to an ensemble (and they note that this result also applies to bagging).
In this work, bagging is evaluated on Auto Imports and Car Evaluation Databases in conjunction with radial basis function and support vector machine as the base learners. The performance of the proposed bagged RBF and SVM classifier is examined in comparison with standalone RBF and SVM.

III. EXISTING CLASSIFICATION METHODS

A. Radial Basis Function

The RBF [13] design involves deciding on their centers and the sharpness (standard deviation) of their Gaussians. Generally, the centres and SD (standard deviations) are decided first by examining the vectors in the training data. RBF networks are trained in a similar way as MLP. The output layer weights are trained using the delta rule. The RBF networks used here may be defined as follows.

1) RBF networks have three layers of nodes: input layer, hidden layer, and output layer.
2) Feed-forward connections exist between input and hidden layers, between input and output layers (shortcut connections), and between hidden and output layers. Additionally, there are connections between a bias node and each output node. A scalar weight is associated with the connection between nodes.
3) The activation of each input node (fanout) is equal to its external input where is the th element of the external input vector (pattern) of the network (denotes the number of the pattern).
4) Each hidden node (neuron) determines the Euclidean distance between "its own" weight vector and the activations of the input nodes, i.e., the external input vector the distance is used as an input of a radial basis function in order to determine the activation of node. Here, Gaussian functions are employed. The parameter of node is the radius of the basis function; the vector is its center.
5) Each output node (neuron) computes its activation as a weighted sum The external output vector of the network, consists of the activations of output nodes, i.e., The activation of a hidden node is high if the current input vector of the network is "similar" (depending on the value of the radius) to the center of its basis function. The center of a basis function can, therefore, be regarded as a prototype of a hyper spherical cluster in the input space of the network. The radius of the cluster is given by the value of the radius parameter.

B. Support Vector Machine

Support vector machines ([4], [3]) are powerful tools for data classification. Classification is achieved by a linear or nonlinear separating surface in the input space of the dataset. The separating surface depends only on a subset of the original data. This subset of data, which is all that is needed to generate the separating surface, constitutes the set of support vectors. In this study, a method is given for selecting as small a set of support vectors as possible which completely determines a separating plane classifier. In nonlinear classification problems, SVM tries to place a linear boundary between two different classes and adjust it in such a way that the margin is maximized [15]. Moreover, in the case of linearly separable data, the method is to find the most suitable one among the hyperplanes that minimize the training error. After that, the boundary is adjusted such that the distance between the boundary and the nearest data points in each class is maximal.

IV. PROPOSED BAGGED ENSEMBLE CLASSIFIERS

Given a set D, of d tuples, bagging works as follows. For iteration i (i =1, 2,…..k), a training set, D_i, of d tuples is sampled with replacement from the original set of tuples, D. The bootstrap sample D_i, by sampling D with replacement, from the given training data set D repeatedly. Each example in the given training set D may appear repeated times or not at all in any particular replicate training data set D_i. A classifier model, M_i, is learned for each training set, D_i. To classify an unknown tuple, X, each classifier, M_i, returns its class prediction, which counts as one vote. The bagged (RBF, SVM), M^*, counts the votes and assigns the class with the most votes to X.

Algorithm: Bagged ensemble classifiers using bagging
Input:

- D, a set of d tuples.
- k = 1, the number of models in the ensemble.
- Base Classifier (Radial Basis Function, Support Vector Machine)

Output: A Bagged (RBF, SVM), M^*
Method:

1. for i = 1 to k do // create k models
2. Create a bootstrap sample, D_i, by sampling D with replacement, from the given training data set D repeatedly. Each example in the given training set D may appear repeated times or not at all in any particular replicate training data set D_i
3. Use D_i to derive a model, M_i;
4. Classify each example d in training data D_i and initialized the weight, W_i for the model, M_i, based on the accuracies of percentage of correctly classified example in training data D_i.
5. endfor

To use the bagged ensemble models on a tuple, X:

1. if classification then
2. let each of the k models classify X and return the majority vote;
3. if prediction then
4. let each of the k models predict a value for X and return the average predicted value;

V. PERFORMANCE EVALUATION MEASURES

A. Cross Validation Technique

Cross-validation [9] sometimes called rotation estimation, is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. 10-fold cross validation is commonly used. In stratified K-fold cross-validation, the folds are selected so that the mean response value is approximately equal in all the folds.

B. Criteria for Evaluation

The primary metric for evaluating classifier performance is classification Accuracy: the percentage of test samples that are correctly classified. The accuracy of a classifier refers to the ability of a given classifier to correctly predict the label of new or previously unseen data (i.e. tuples without class label information). Similarly, the accuracy of a predictor refers to how well a given predictor can guess the value of the predicted attribute for new or previously unseen data.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

A. Auto Imports Data base Description

This data set consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) its assigned insurance risk rating, (c) its normalized losses in use as compared to other cars. The second rating corresponds to the degree to which the auto is more risky than its price indicates. Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuarians call this process "symboling". A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.
The third factor is the relative average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification (two-door small, station wagons, sports/speciality, etc...), and represents the average loss per car per year.

TABLE I: PROPERTIES OF AUTO IMPORTS DATABASE

Data Set Characteristics:	Multivariate	**Number of Instances:**	205
Attribute Characteristics:	Categorical, Integer, Real	**Number of Attributes:**	26
Associated Tasks:	Regression	**Missing Values**	Yes

It contains the following attributes:

1. symboling: -3, -2, -1, 0, 1, 2, 3.
2. normalized-losses: continuous from 65 to 256.
3. make: alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, Volvo
4. fuel-type: diesel, gas.
5. aspiration: std, turbo.
6. num-of-doors: four, two.
7. body-style: hardtop, wagon, sedan, hatchback, convertible.
8. drive-wheels: 4wd, fwd, rwd.
9. engine-location: front, rear.
10. wheel-base: continuous from 86.6 120.9.
11. length: continuous from 141.1 to 208.1.
12. width: continuous from 60.3 to 72.3.
13. height: continuous from 47.8 to 59.8.
14. curb-weight: continuous from 1488 to 4066.
15. engine-type: dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16. num-of-cylinders: eight, five, four, six, three, twelve, two.
17. engine-size: continuous from 61 to 326.
18. fuel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. bore: continuous from 2.54 to 3.94.
20. stroke: continuous from 2.07 to 4.17.
21. compression-ratio: continuous from 7 to 23.
22. horsepower: continuous from 48 to 288.
23. peak-rpm: continuous from 4150 to 6600.
24. city-mpg: continuous from 13 to 49.
25. highway-mpg: continuous from 16 to 54.
26. price: continuous from 5118 to 45400.

B. Car Evaluation Database Description

The dataset is obtained from UCI Machine Learning Repository, which is supplied by the University of California. The car evaluation database was originally derived from a simple hierarchical decision model. The model evaluates cars according to the following concept structure:

CAR - Car acceptability
PRICE - Overall price
Buying - Buying price
Maint - Price of maintenance
TECH - Technical characteristics
COMFORT - Level of comfort
Doors - Number of doors
Persons - Capacity in terms of passengers
Lug_boot - The size of luggage boot
Safety - Estimated safety of the car

TABLE III
PROPERTIES OF CAR EVALUATION DATABASE

Data Set Characteristics:	Multivariate	**Number of Instances:**	1728
Attribute Characteristics:	Categorical	**Number of Attributes:**	6
Associated Tasks:	Classification	**Missing Values**	No

PRICE, TECH, and COMFORT are three immediate concepts. Every concept is related to its lower level descendants by a set of examples. The car evaluation database contains examples with the structural information removed, i.e., directly relates CAR to six input attributes: buying, maint, doors, persons, lug_boot, and safety. There are 1,728 instances that completely cover the attribute space with 6 attributes (no missing attribute values) as follows:

• buying: v-high, high, med, low
• maint: v-high, high, med, low
• doors: 2, 3, 4, 5-more
• persons: 2, 4, more
• lug_boot: small, med, big
• safety low, med, high

The class distribution, which is the number of instances per class is shown in Table III

TABLE IIIII
CLASS DISTRIBUTION

Class Name	Number of instance per class	Percentage (%)
Unaac	1210	70.023
Acc	384	22.222
Good	69	3.993
Vgood	65	3.762

C. Experiments and Analysis

1) Auto Imports Database: The auto imports database is taken to evaluate the proposed bagged RBF and SVM for automobile prediction system.

TABLE IV
THE PERFORMANCE OF EXISTING AND PROPOSED BAGGED CLASSIFIERS FOR AUTIO IMPORT DATABASE

Dataset	Classifiers	Classification Accuracy
Auto Imports Database	Existing RBF Classifier	61.95 %
	Proposed Bagged RBF Classifier	87.80 %
	Existing SVM Classifier	71.21 %
	Proposed Bagged SVM	89.26 %

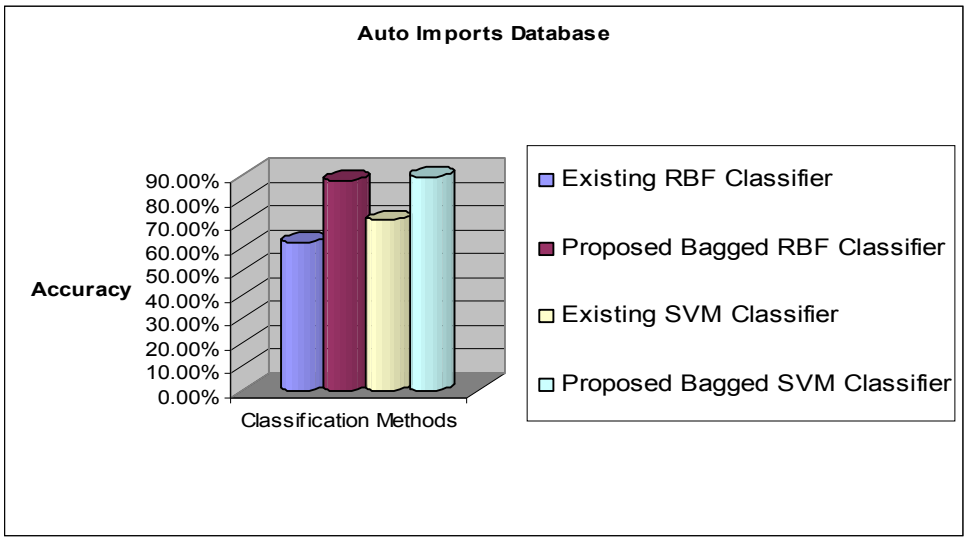


Fig. 1 Classification Accuracy of Existing and Proposed Bagged Classifiers using Auto Imports Database

2) Car Evaluation Database: The car evaluation database is taken to evaluate the proposed bagged SVM and RBF for car marketing prediction system.

TABLE V
CLASSIFICATION ACCURACY OF EXISTING AND PROPOSED BAGGED CLASSIFIERS FOR CAR EVALUATION DATABASE

Dataset	**Classifiers**	**Classification Accuracy**
Car Evaluation Database	Existing RBF Classifier	88.25 %
	Proposed Bagged RBF Classifier	93.86 %
	Existing SVM Classifier	93.75 %
	Proposed Bagged SVM	95.48 %

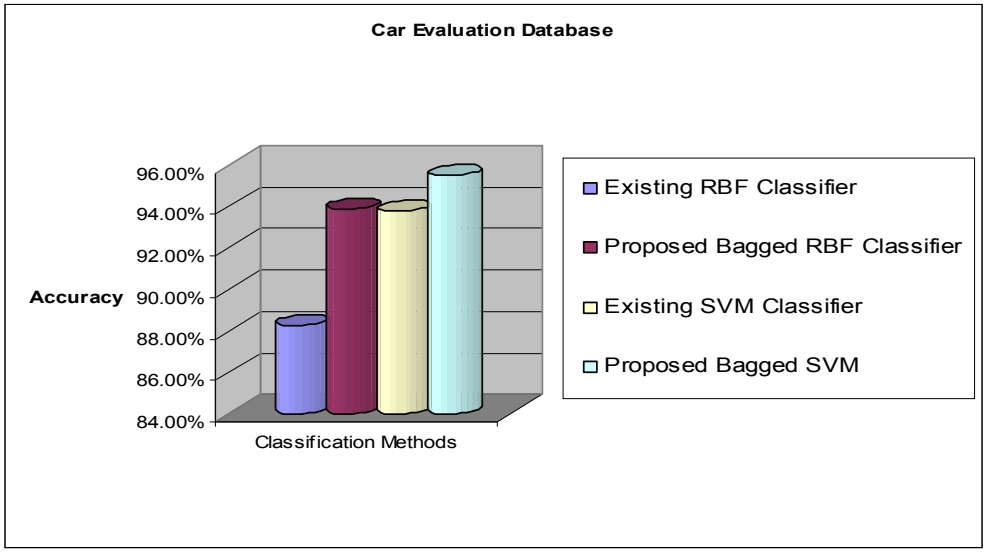


Fig. 2 Classification Accuracy of Existing and Proposed Bagged Classifiers using Car Evaluation Database

In this research work, new ensemble classification method is proposed using bagging classifier in conjunction with support vector machine as the base learner and the performance is analyzed in terms of accuracy. Here, the base classifiers are constructed using radial basis function and support vector machine. 10-fold cross validation [11] technique is applied to the base classifiers and evaluated classification accuracy. Bagging is performed with radial basis function and support vector machine to obtain a very good classification performance. Table IV and V shows classification performance for real and benchmark datasets of intrusion detection, direct marketing, signature verification using existing and proposed bagged radial basis function and support vector machine. The analysis of results shows that the proposed bagged radial basis function and support vector machine are shown to be superior to individual approaches for automobile data in terms of classification accuracy. According to Fig. 1 and 2 proposed combined model show significantly larger improvement of classification accuracy than the base classifiers. This means that the combined method is more accurate than the individual methods for the automobile data.

The χ^2 statistic is determined for the above approach and the critical value is found to be less than 0.455. Hence corresponding probability is $p < 0.5$. This is smaller than the conventionally accepted significance level of 0.05 or 5%. Thus examining a χ^2 significance table, it is found that this value is significant with a degree of freedom of 1. In general, the result of χ^2 statistic analysis shows that the proposed classifier is significant at $p < 0.05$ than the existing classifier.

VII. CONCLUSIONS

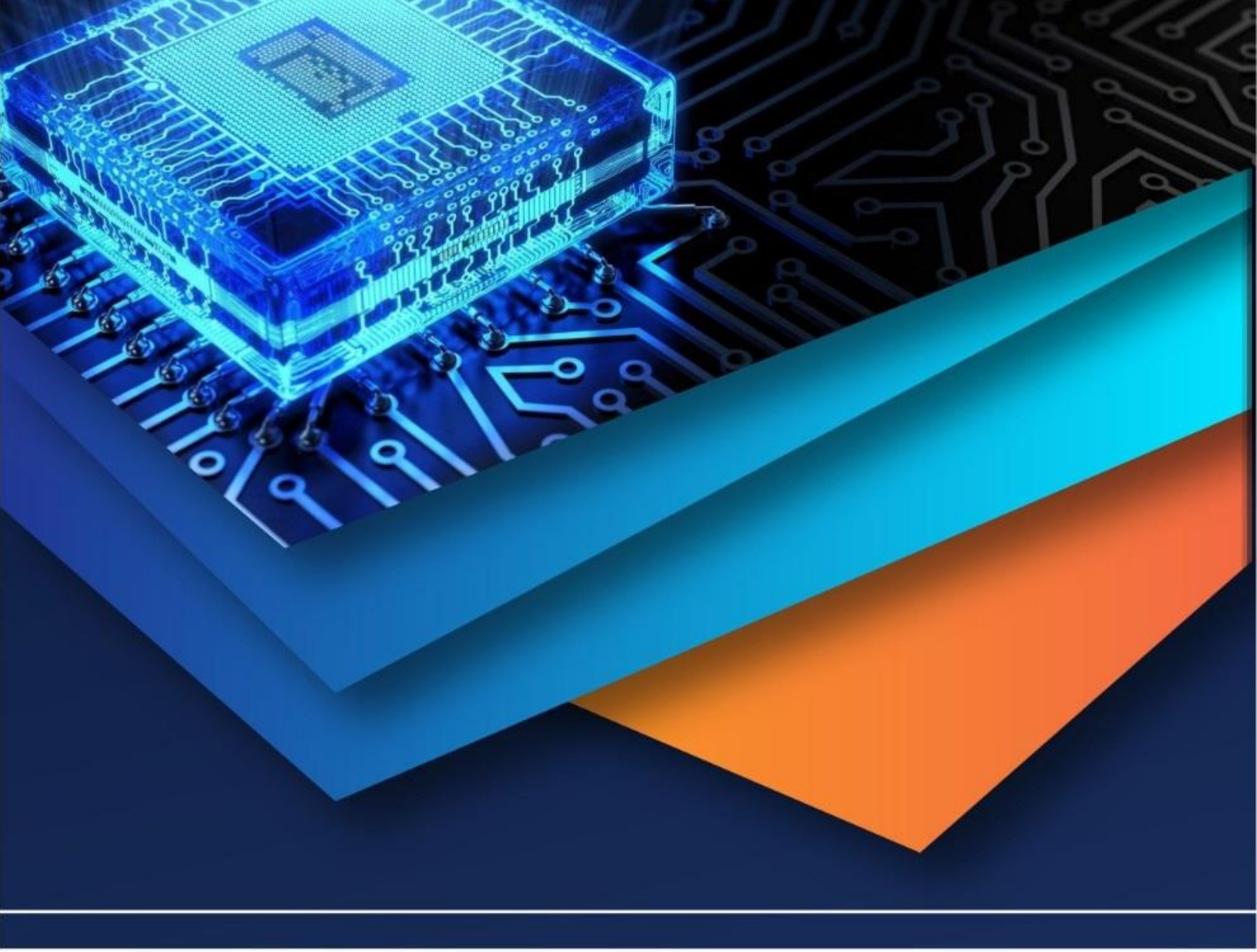
In this research work, new combined classification method is proposed using bagging classifier in conjunction with radial basis function and support vector machine as the base learner and the performance comparison has been demonstrated using Auto Imports and Car Evaluation Databases in terms of accuracy. This research has clearly shown the importance of using ensemble approach for automobile data like Auto Imports and Car Evaluation Databases. An ensemble helps to indirectly combine the synergistic and complementary features of the different learning paradigms without any complex hybridization. Since all the considered performance measures could be optimized, such systems could be helpful in several real world automobile data. The high classification accuracy has been achieved for the ensemble classifier compared to that of single classifier. The proposed bagged radial basis function and support vector machine is shown to be significantly higher improvement of classification accuracy than the base classifiers. The real dataset of automobile could be detected with high accuracy for homogeneous model. The future research will be directed towards developing more accurate base classifier particularly for the automobile data.

VIII. ACKNOWLEDGMENT

Author gratefully acknowledges the authorities of Annamalai University for the facilities offered and encouragement to carry out this work.

REFERENCES

[1] Breiman, L,"Bagging predictors", Machine Learning, 24 (2), 123-140, 1996a.
[2] Breiman, L,"Stacked Regressions", Machine Learning, 24(1), pp.49-64,1996c.
[3] Burges, C. J. C. (1998), "A tutorial on support vector machines for pattern recognition", Data Mining and Knowledge Discovery, 2(2), pp.121-167.
[4] Cherkassky, V. and Mulier, F, "Learning from Data - Concepts, Theory and Methods", John Wiley & Sons, New York, 1998.
[5] N. Cristianini, B. Schoelkopf, "Support vector machines and kernel methods, the new generation of learning machines". Artificial Intelligence Magazine, 23(3), pp. 31–41, 2002.
[6] D. Delen, G. Walker, and A. Kadam, "Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods", Artificial Intelligence in Medicine, Elsevier, pp. 121-130, 2004.
[7] Freund, Y., & Schapire, R. E, "Experiments with a new boosting algorithm", In Proc. 13th International Conference on Machine Learning, pp. 148-146, Morgan Kaufmann, 1996.
[8] T. Ho, J. Hull, S. Srihari, "Decision combination in multiple classifier systems", IEEE Transactions on Pattern Analysis and Machine Intelligence, 16, pp. 66–75, 1994.
[9] Jiawei Han, Micheline Kamber, "Data Mining – Concepts and Techniques", Elsevier Publications, 2003.
[10] J. Kittler, (1998), "Combining classifiers: a theoretical framework", Pattern Analysis and Applications, 1, pp.18–27.
[11] Kohavi, R, "A study of cross-validation and bootstrap for accuracy estimation and model selection", Proceedings of International Joint Conference on Artificial Intelligence, pp.1137–1143, 1995.
[12] J. A. Marchant and C. M. Onyango, "Comparison of a Bayesian Classifier with a Multilayer Feed-Forward Neural Network using the Example of Plant/Weed/Soil Discrimination", Computers and Electronics in Agriculture, Elsevier, 39: pp. 3-22, 2003.
[13] Oliver Buchtala, Manuel Klimek, and Bernhard Sick, Member, IEEE, "Evolutionary Optimization of Radial Basis Function Classifiers for Data Mining Applications", IEEE Transactions on systems, man, and cybernetics—part b: cybernetics, 35(5), 2005.
[14] Schapire, R., Freund, Y., Bartlett, P., and Lee, W, "Boosting the margin: A new explanation for the effectives of voting methods", In proceedings of the fourteenth International Conference on Machine Learning, Nashville, TN, pp. 322-330, 1997.
[15] Vanajakshi, L. and Rilett, L.R, "A Comparison of the Performance of Artificial Neural Network and Support Vector Machines for the Prediction of Traffic Speed", IEEE Intelligent Vehicles Symposium, University of Parma, Parma, Italy, IEEE, pp.194-199, 2004.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY