Classifier Based Ensembles for Car Evaluation Dataset

M.Govindarajan
Assistant Professor
Department of Computer Science and Engineering,
Annamalai University, Annamalai Nagar – 608002, Tamil Nadu, India

Abstract : In this research work, new ensemble classification methods are proposed with homogeneous ensemble classifier using bagging and heterogeneous ensemble classifier using arcing and their performances are analyzed in terms of accuracy. A Classifier ensemble is designed using Radial Basis Function (RBF) and Support Vector Machine (SVM) as base classifiers. The feasibility and the benefits of the proposed approaches are demonstrated by the means of car evaluation dataset. The main originality of the proposed approach is based on three main parts: pre-processing phase, classification phase and combining phase. A wide range of comparative experiments are conducted for car evaluation dataset. The proposed ensemble methods provide significant improvement of accuracy compared to individual classifiers. Also heterogeneous models exhibit better results than homogeneous models for car evaluation dataset.

IndexTerms - **Accuracy, Arcing, Bagging, Ensemble, Radial Basis Function, Support Vector Machine**.

I. INTRODUCTION

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.

One of the major developments in machine learning in the past decade is the ensemble method, which finds highly accurate classifier by combining many moderately accurate component classifiers. This paper proposes new ensemble classification methods to improve the classification accuracy. The main purpose of this paper is to apply homogeneous and heterogeneous ensemble classifiers for car evaluation dataset to improve classification accuracy.

Organization of this paper is as follows. Section 2 describes the related work. Section 3 presents proposed methodology 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

In the field of automobile lot of research has been done in which many techniques are covered and still many remains to be covered.

M.Govindarajan and A.Mishra (2014a) have been investigated for automobile data and evaluated their performance based on classification accuracy. RBF and SVM have been explored as hybrid models. Next a hybrid RBF-SVM model and RBF, SVM models as base classifiers are designed. Finally, hybrid systems are proposed to make optimum use of the best performances delivered by the individual base classifiers and the hybrid approach.

M.Govindarajan and A.Mishra (2014b) 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.

Yingju Xia et al. (2015) proposed a novel feature reduction method which adopts ensemble approach to measure the divergence between the training set and test set and use the divergence to supervise the feature reduction procedure. The proposed method uses pairwise measure to get the diversity between classifiers and selects the complementary classifiers to get the pseudo labels on test set. The pseudo labels are used to measure the divergence between training set and test set. The feature reduction algorithm merges the adjacent feature space according to the divergence, such reduce the feature number.

Uma R. Salunkhea and Suresh N. Mali (2016) presented a novel approach that initially applies pre-processing to the imbalanced dataset in order to reduce the imbalance between the classes. The pre-processed data is provided as training dataset to the classifier ensemble that introduces diversity by using different training datasets as well as different classifier models.

Wei Liu et al. (2017) proposed a method, which integrates deep neural networks with balanced sampling in this paper. The proposed method consists of two main stages. In the first stage, data augmentation with balanced sampling is applied to alleviate the unbalanced data set problem. In the second stage, an ensemble of convolutional neural network models with different architectures is constructed with parameters learned on the augmented training data set.

Ayşe Eldem (2018) aimed to find the most appropriate class information according to the pre-defined class information of the data compared to other data mining techniques, clustering and association rules. A correctly trained model will provide more accurate classification of new data. In this study, many classification methods such as Decision Trees, Generalized Linear Model, Naive Bayes, Random Forest have been used in the classification of car dataset and performance comparison of these methods has been made.

Pelin YJldJrJm et al. (2019) proposed a novel ensemble-based ordinal classification (EBOC) approach which suggests bagging and boosting (AdaBoost algorithm) methods as a solution for ordinal classification problem in transportation sector. This article also compares the proposed EBOC approach with ordinal class classifier and traditional tree-based classification algorithms (i.e., C4.5 decision tree, RandomTree, and REPTree) in terms of accuracy. The results indicate that the proposed EBOC approach achieves better classification performance than the conventional solutions.

In this paper, a hybrid system is proposed using radial basis function and support vector machine and the effectiveness of the proposed bagged RBF, bagged SVM and RBF-SVM hybrid system is evaluated by conducting several experiments on car evaluation dataset. The performance of the proposed bagged RBF, bagged SVM, and RBF-SVM hybrid classifiers are examined in comparison with standalone RBF and standalone SVM classifier and also heterogeneous models exhibits better results than homogeneous models for car evaluation dataset.

III. PROPOSED METHODOLOGY

3.1Preprocessing

Before performing any classification method the data has to be preprocessed. In the data preprocessing stage it has been observed that the datasets consist of many missing value attributes. By eliminating the missing attribute records may lead to misclassification because the dropped records may contain some useful pattern for Classification. The dataset is preprocessed by removing missing values using supervised filters.

3.2 Existing Classification Methods

3.2.1 Radial Basis Function Neural Network

The Radial Basis Function Network (RBF) is in its simplest form a three layered feed forward neural network with one input layer, one hidden layer and one output layer (R. Callan, 1998). It differs from an MLP in the way the hidden layer performs its computation. The connection between the input layer and the output layer is nonlinear, while the connection between the hidden layer and the output layer is linear. RBF networks are instance based, meaning that it will compare and evaluate each training case to the previous examined training cases. In an MLP all instances are evaluated once while in an RBF network the instances are evaluated locally (Tom M, 1997). Instance based methods use nearest neighbor and locally weighted regression methods. An RBF network can be trained more efficiently than a neural net using backpropagation since the input and output layer are trained separately.

3.2.2 Support Vector Machine

Support Vector Machines has been introduced by Vapnik and his colleagues (C. Cortes and V. Vapnik, 1995), SVM models are very similar to classical multilayer perceptron neural networks used for classification (R. Hua, Dai liankui, 2010), but recently they have been extended to solve regression problems (V. Vapnik et al., 1997). SVM is very similar to an ANN since both receive input data and provide output data. For regression, the input and output of SVM are identical to the ANN. However, what makes the SVM primarily better is that the SVM does not suffer from over fitting like ANN does. So, the ANN memorizes the input data on the training stage and will not perform well at the testing data.

3.3 Homogeneous Ensemble Classifiers

3.3.1 Proposed Bagged RBF and SVM Classifiers

Given a set D, of d tuples, bagging (Breiman, L. 1996a) 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, created by sampling D with replacement, from the given training data set D repeatedly. Each example in the given training set D may appear repeatedly 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 and SVM, M^*, counts the votes and assigns the class with the most votes to X.

Algorithm: RBF and SVM ensemble classifiers using bagging
Input:

- D, a set of d tuples.
- k = 2, the number of models in the ensemble.
- Base Classifiers (Radial Basis Function, Support Vector Machine)

Output: Bagged RBF and 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 RBF and SVM 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;

3.4 Heterogeneous Ensemble Classifiers

3.4.1 Proposed RBF-SVM Hybrid System

Given a set D, of d tuples, arcing (Breiman. L, 1996) 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. some of the examples from the dataset D will occur more than once in the training dataset D_i. The examples that did not make it into the training dataset end up forming the test dataset. Then a classifier model, M_i, is learned for each training examples d from training dataset 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 hybrid classifier (RBF-SVM), M^*, counts the votes and assigns the class with the most votes to X.

Algorithm: Hybrid RBF-SVM using Arcing Classifier
Input:

- D, a set of d tuples.
- k = 2, the number of models in the ensemble.
- Base Classifiers (Radial Basis Function, Support Vector Machine)

Output: Hybrid RBF-SVM model, M^*.
Procedure:

1. For i = 1 to k do // Create k models
2. Create a new training dataset, D_i, by sampling D with replacement. Same example from given dataset D may occur more than once in the training dataset 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 hybrid model 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;

The basic idea in Arcing is like bagging, but some of the original tuples of D may not be included in Di, where as others may occur more than once.

IV. PERFORMANCE EVALUATION MEASURES

4.1 Cross Validation Technique

Cross-validation (Jiawei Han and Micheline Kamber**,** 2003) 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.

4.2 Criteria for Evaluation

The primary metric for evaluating classifier performance is classification Accuracy: the percentage of test samples that 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.

V. EXPERIMENTAL RESULTS AND DISCUSSION

5.1 Car Evaluation dataset 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

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.

5.2 Experiments and Analysis

In this section, new ensemble classification methods are proposed using classifiers in both homogeneous ensembles using bagging and heterogeneous ensembles using arcing classifier and their performances are analyzed in terms of accuracy.

Table 5.1: Performance Comparison of the Homogeneous and Heterogeneous Ensemble Classifiers for Car Evaluation dataset

Dataset	Classifiers	Classification Accuracy
Car Evaluation	RBF	88.25 %
	Proposed Bagged RBF	93.86 %
	SVM	93.75 %
	Proposed Bagged SVM	95.48 %
	Proposed Hybrid RBF-SVM	98.66 %

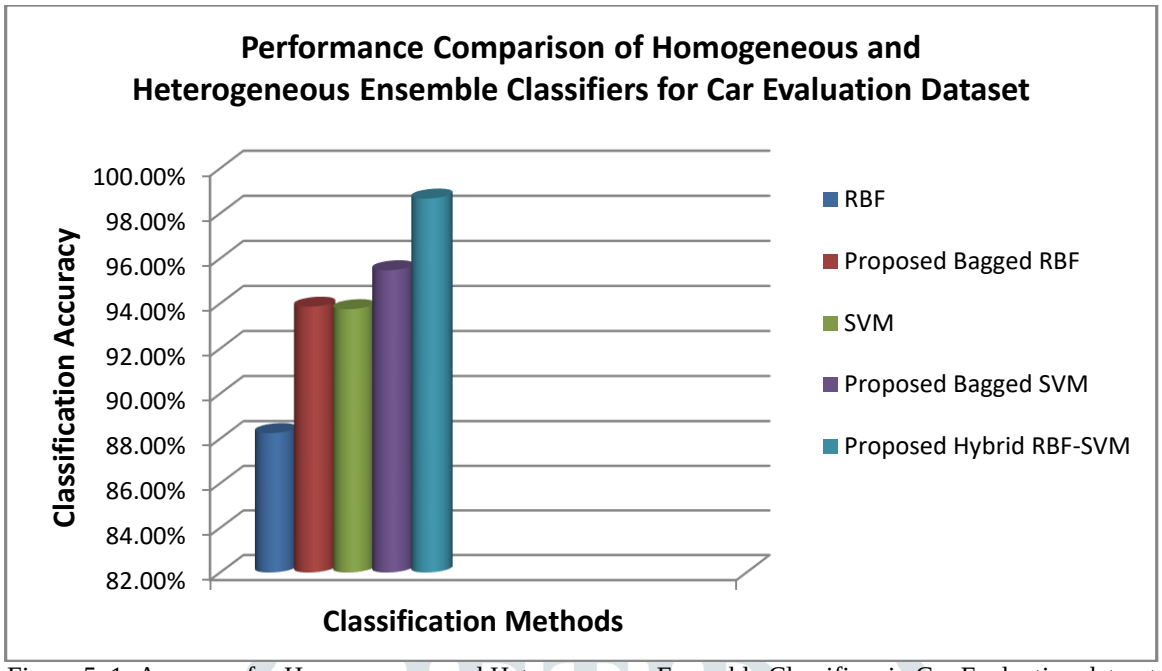

Figure 5. 1: Accuracy for Homogeneous and Heterogeneous Ensemble Classifiers in Car Evaluation dataset

A wide range of comparative experiments are conducted for car evaluation dataset. The accuracy of base classifiers is compared with homogeneous and heterogeneous models for car evaluation dataset as given in Table 5.1. According to Figure 5.1, the proposed hybrid methods provide significant improvement of accuracy compared to individual classifiers and also heterogeneous models exhibits better results than homogeneous models.

VI. CONCLUSION

In this research work, new combined classification methods are proposed using classifiers in homogeneous ensembles using bagging and the performance comparisons have been demonstrated using car evaluation dataset in terms of accuracy. Here, the proposed bagged radial basis function and bagged support vector machine combines the complementary features of the base classifiers. Similarly, new hybrid RBF-SVM models are designed in heterogeneous ensembles involving RBF and SVM models as base classifiers and their performances are analyzed in terms of accuracy.
The experiment results lead to the following observations.

- SVM exhibits better performance than RBF in the important respects of accuracy.
- The proposed bagged methods are shown to be significantly higher improvement of classification accuracy than the base classifiers.
- The hybrid RBF-SVM shows higher percentage of classification accuracy than the base classifiers.
- The proposed ensemble methods provide significant improvement of accuracy compared to individual classifiers.
- The heterogeneous models exhibit better results than homogeneous models for car evaluation dataset.
- Assessment of performance is based on the calculation of the χ^2 statistic for all the approaches and their critical values are found to be less than 0.455. Hence their 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 classifiers are significant at $p < 0.05$ than the existing classifiers.
- The future research will be directed towards developing more accurate base classifiers particularly for the car evaluation dataset.

VII. ACKNOWLEDGMENT

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

REFERENCES

[1] Ayşe Eldem. 2018. Comparison of Prediction Success Performances for Classification Methods. International Journal of Engineering Science Invention, 7(12), Ver. II: 63-65.
[2] Breiman. L. 1996. Bias, Variance, and Arcing Classifiers. Technical Report 460, Department of Statistics, University of California, Berkeley, CA.
[3] Breiman, L. 1996a. Bagging predictors. Machine Learning, 24(2):123–140.
[4] Callan, R. 1998. Essence of neural networks. Prentice Hall PTR Upper Saddle River, NJ, USA.
[5] Cortes, C. and Vapnik, V. 1995. Support vector networks, Machine learning, 20(3):273-297.
[6] Govindarajan, M. and Mishra, A. 2014a. Development of a hybrid ensemble approach for Automobile data. International Journal of Engineering Sciences & Research Technology, 3(12):387-392.

[7] Govindarajan, M. and Mishra, A. 2014b. Performance Analysis of Automobile Data using Bagged Ensemble Classifiers. International Journal for Research in Applied Science & Engineering Technology, 2(XII):257-263.
[8] Hua, R. and Dai liankui. 2010. support vector machine classification and regression based hybrid modeling method and its application in raman spectral analysis. Chinese Journal of Scientific Instrument, (11):2440-2446.
[9] Jiawei Han, Micheline Kamber. 2003. Data Mining – Concepts and Techniques, Elsevier Publications.
[10] Pelin YJldJrJm, UlaG K. Birant, and Derya Birant. 2019. EBOC: Ensemble-Based Ordinal Classification in Transportation. Journal of Advanced Transportation, Volume 2019:1-17.
[11] Tom M. Mitchell. 1997. Machine Learning, McGraw-Hill, New York.
[12] Uma R. Salunkhea, Suresh N. Mali. 2016. Classifier Ensemble Design for Imbalanced Data Classification: A Hybrid Approach. Procedia Computer Science, 85:725 – 732.
[13] Vapnik, V. Golowich, S. and Smola, A. 1997. Support vector method for function approximation, regression estimation, and signal processing. Advances in neural information processing systems, 281-287.
[14] Wei Liu, Miaohui Zhang, Zhiming Luo, and Yuanzheng Cai. 2017. An Ensemble Deep Learning Method for Vehicle Type Classification on Visual Traffic Surveillance Sensors. IEEE Access: Special Section On Visual Surveillance And Biometrics: Practices, Challenges, And Possibilities,5:24417-24425.
[15] Yingju Xia, Cuiqin Hou, Zhuoran Xu, Jun Sun. 2015. Feature Reduction Using Ensemble Approach. 29th Pacific Asia Conference on Language, Information and Computation: Posters, Shanghai, China, October 30-November 1, 2015:309-318.