Classification of Imbalanced Data Using Heterogeneous Ensemble Model

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

Abstract- Multiclass problem has continued to be an active research area due to the challenges paused by the issue of imbalance datasets and lack of a unifying classification algorithms. In this research work, new ensemble classification method is proposed with 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 standard dataset of automobile. The main originality of the proposed approach is based on three main parts: pre-processing phase, classification phase and combining phase. Wide ranges of comparative experiments are conducted for standard dataset of automobile. The performance of the proposed heterogeneous ensemble classifier is compared to the performance of other standard heterogeneous ensemble methods. The standard heterogeneous ensemble methods include majority voting, stacking. The proposed ensemble method provides significant improvement of accuracy compared to individual classifiers and also performs significantly better than majority voting and stacking.

Keywords – Accuracy, Arcing, Ensemble, Radial Basis Function, Support Vector Machine

I. INTRODUCTION

In data mining, the classification task encompasses constructing prediction models and utilizing them to predict categories (class labels) for new unseen observations. Generatingpredictionmodelsareachievedbyutilizingclassification algorithms; many options are available for thispurpose. According to the literature, there is no single “super”classificationalgorithmthatproducesthebestresultsfor all cases (datasets) [8]. Consequently, several researchersattemptedtoimprovetheeffectivenessofclassificationbyutilizing a group of classifiers instead of using a single classifier [1]. Utilizing a collection of classifiers for predicting classlabels is referred to as the “ensemble” classification [8]. Integratingseveralclassificationmodelsisconsideredanattractiveresearch topic, due to the potential performance improvementover a single classification.

Ensemble classification has twomain categories: (i) Homogenous ensemble and (ii) Heterogeneous ensemble. In Homogenous ensembles, all classifiersforming the ensemble are generated using the same classification algorithm; Bagging is an example of this category whereall base classifiers forming the ensemble are produced usingtheDecisionTreealgorithm.RegardingHeterogeneousensemble,theclassifiersformingtheensembleareproducedthrough utilizing several classification algorithms [2]. An example ofheterogeneous ensemble is the model constructed by [5], where Support Vector Machines, Artificial NeuralNetworks, Memory-Based Learning, Decision Trees, BaggedDecision Trees, Boosted Decision Trees and Boosted Stumpsalgorithms were used to construct an ensemble classificationmodel.

This paper proposes new ensemble classification method to improve the classification accuracy. The main purpose of this paper is to apply heterogeneous ensemble classifiers for standard dataset of automobile 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.

Reference [11] proposed a novel RE-sample and Cost-Sensitive Stacked Generalization (RECSG) method based on 2-layer learning models. The first step is Level0 model generalization including data pre-processing and base model training. The second step is Level1 model generalization involving cost-sensitive classifier and logistic regression algorithm. In the learning phase, pre-processing techniques can be embedded in imbalance data learning methods. In the cost-sensitive algorithm, cost matrix is combined with both data characters and algorithms. In the RECSG method, ensemble algorithm is combined with imbalance data techniques.

Reference [14] 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.

Reference [13] proposed a hybrid optimal ensemble classifier framework that combines density-based undersampling and cost-effective methods through exploring state-of-the-art solutions using multi-objective optimization algorithm. Specifically, a density-based undersampling method is firstlydeveloped to select informative samples from the original training data with probability-based data transformation, which enables to obtain multiple subsets following a balanced distribution across classes. Second, the cost-sensitive classification method is exploited to address the incompleteness of information problem via modifying weights of misclassified minority samples rather than the majority ones. Finally, introduced a multi-objective optimization procedure and utilize connections between samples to self-modify the classification result using an ensemble classifier framework.

Reference [7] proposed trimming approach finds an optimal subset of classifiers to form the desired heterogeneous ensemble. The main challenge is how to detect poor performance classifiers. To address this issue, the differences in effectiveness between base classifiers forming the ensemble are utilized to spot weak classifiers.

Reference [10] proposed a density-based random forest algorithm (DBRF) to improve the prediction performance, especially for minority classes. DBRF is designed to recognize boundary samples as the most difficult to classify and then use a density-based method to augment them. Subsequently, two different random forest classifiers wereconstructed to model the augmented boundary samples and the original dataset dependently, and the final output was determined using a bagging technique. The performance of the proposed hybrid RBF-SVM classifier is examined in comparison with standalone RBF and standalone SVM classifier and also standard heterogeneous models for standard dataset of automobile.

III. Proposed Methodology

A. Preprocessing

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.\

B. Existing Classification Methods

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 [4]. 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 [15]. 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.

2) Support Vector Machine

Support Vector Machines has been introduced by Vapnik and his colleagues [6], SVM models are very similar to classical multilayer perceptron neural networks used for classification [9], but recently they have been extended to solve regression problems [16]. 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.

C. Heterogeneous Ensemble Classifiers

1) Weighted Majority Algorithm

In machine learning, Weighted Majority Algorithm (WMA) is a meta-learning algorithm used to construct a compound algorithm from a pool of prediction algorithms, which could be any type of learning algorithms, classifiers, or even real human experts. The algorithm assumes that there might not be prior knowledge about the accuracy of the algorithms in the pool, but there are sufficient reasons to believe that one or more will perform well.

2) Stacking

Stacking (sometimes called stacked generalization) involves training a learning algorithm to combine the predictions of several other learning algorithms. First, all of the other algorithms are trained using the available data, then a combiner algorithm is trained to make a final prediction using all the predictions of the other algorithms as additional inputs. Stacking typically yields performance better than any single one of the trained models. It has been successfully used on both supervised learning tasks (regression) and unsupervised learning (density estimation).

3) Proposed RBF-SVM Hybrid System

Given a set D, of d tuples, arcing [3] 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 ArcingClassifier

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 din 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

A. Cross Validation Technique

Cross-validation [12] 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 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

A. Vehicle dataset Description

This dataset classifies a given silhouette from four different vehicle types, with a set of features that are extracted from the silhouette by the Hierarchical Image Processing System extension BINATTS.

B. Experiments and Analysis

In this section, new ensemble classification method is proposed with heterogeneous ensemble using arcing and their performances are analyzed in terms of accuracy.

Table - 1The Performance of Base Classifiers and Heterogeneous Ensemble Classifiers for Automobile

Dataset	**Classifiers**	**Classification Accuracy**
Vehicle	**RBF**	**66.66 %**
	SVM	**74.34 %**
	Proposed Hybrid RBF-SVM	**80.26 %**
	Voted RBF-SVM	**71.27 %**
	Stacked RBF-SVM	**69.03 %**

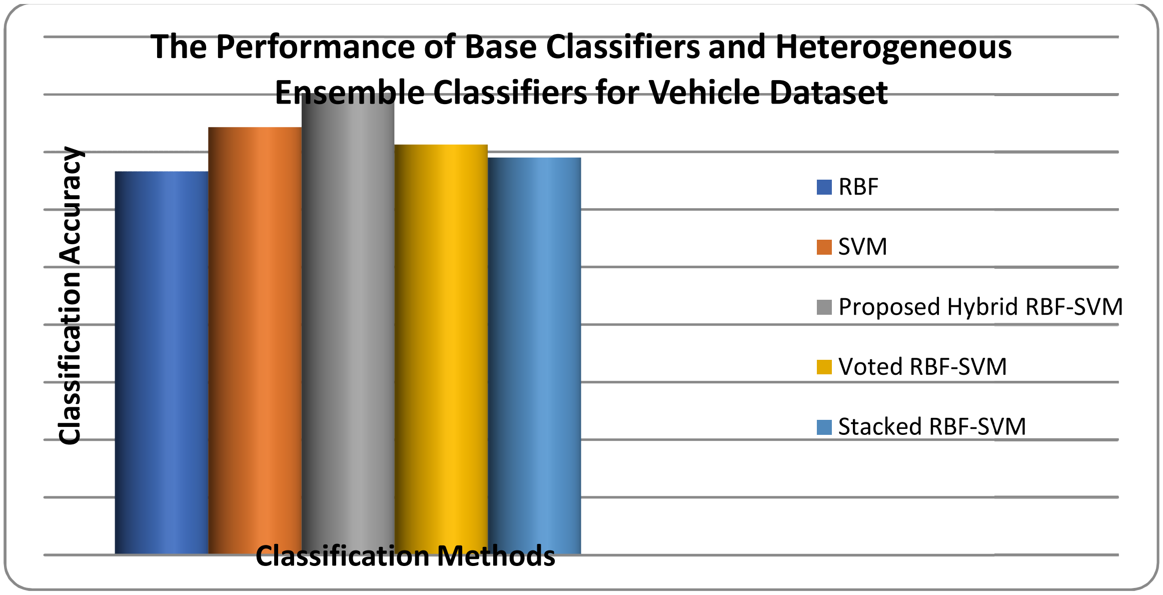

Figure 1. Accuracy for Heterogeneous Ensemble Classifiers in Vehicle dataset

In this research work, new ensemble classification method is proposed with heterogeneous ensemble classifier using arcing and their performances are analyzed in terms of accuracy. Here, the base classifiers are constructed using radial basis function and Support Vector Machine. Arcing is performed with radial basis function classifier and support vector machine to obtain a very good classification performance. The analysis of results shows that the proposed hybrid RBF-SVM classifier is shown to be superior to individual approaches for standard dataset of automobile problem in terms of classification accuracy. According to Figure 1 proposed hybrid model show significantly larger improvement of classification accuracy than the base classifiers and the results are found to be statistically significant. Table 1 compares the performance of proposed hybrid RBF-SVM to the performance of majority voting and stacking with RBF and SVM. The proposed hybrid RBF-SVM performs significantly better than majority voting and stacking on standard dataset of automobile.

VI. CONCLUSION

In this research work, new hybrid classification method is proposed with heterogeneous ensembles using arcing and the performance comparisons have been demonstrated using standard dataset of automobile in terms of accuracy. Here, the proposed hybrid RBF-SVM combines the complementary features of the base classifiers. The performance of the proposed heterogeneous ensemble classifier is compared to the performance of other standard heterogeneous ensemble methods. The standard heterogeneous ensemble methods include majority voting, stacking. The experiment results lead to the following observations.

- SVM exhibits better performance than RBF in the important respects of accuracy.
- The proposed hybrid RBF-SVM method is shown to be significantly higher improvement of classification accuracy than the base classifiers.
- The proposed hybrid RBF-SVM method provide significant improvement of accuracy compared to individual classifiers and the proposed hybrid RBF-SVM performs significantly better than majority voting and stacking.

The future research will be directed towards developing more accurate base classifiers particularly for the automobile problem.

ACKNOWLEDGEMENT

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

REFERENCES

[1] E. Alshdaifat, F. Coenen, K. Dures, “ A directed acyclic graph (DAG)ensemble classification model: An alternative architecture for hierarchical classification”, *Int. J. Data Warehous. Min. (IJDWM),* 13(3), pp.73–90, 2017.

[2] L.Breiman, “Baggingpredictors”,*Mach.Learn.,*24(2), 123–140, 1996.

[3] Breiman, L, “Bias, Variance, and Arcing Classifiers”, *Technical Report 460*, Department of Statistics, University of California, Berkeley, CA, pp.1-23, 1996a.

[4] R. Callan, “Essence of neural networks”, Prentice Hall PTR Upper Saddle River, NJ, USA, 1998.

[5] R.Caruana,A.Niculescu-Mizil,G.Crew,A.Ksikes, “ Ensembleselection from libraries of models”, *Proceedings of the Twenty-firstInternationalConferenceonMachineLearning*, p p.18, 2004.

[6] C. Cortes and V. Vapnik, “Support vector networks”, *Machine learning*, 20(3), pp.273-297, 1995.

[7] Esra’a Alshdaifat, MalakAl-hassan, Ahmad Aloqaily, “Effective heterogeneous ensemble classification: An alternative approach for selecting base classifiers”, *ICT Express*, 7(3), pp. 342-349, 2021.

[8] J.Han,M.Kamber,J.Pei., “DataMiningConceptsandTechniques”,thirded.,MorganKaufmannPublishers,Waltham,Mass, 2012.

[9] R. Hua, Dai liankui, “support vector machine classification and regression based hybrid modeling method and its application in raman spectral analysis*”, Chinese Journal of Scientific Instrument*, 11, pp.2440-2446, 2010.

[10] Jia Dong and Quan Qian, “ A Density-Based Random Forest for Imbalanced Data Classification”, *Future Internet*, 14, 90, pp.1-20, 2022.

[11] Jianhong Yan and Suqing Han, “Classifying Imbalanced Data Sets by a NovelRE-Sample and Cost-Sensitive Stacked Generalization Method*”, Mathematical Problems in Engineering*, Volume 2018, 1-13, 2018.

[12] Jiawei Han, Micheline Kamber, “Data Mining – Concepts and Techniques”, Elsevier Publications, 2003.

[13] Kaixiang Yang, Zhiwen Yu, Senior Member, IEEE, Xin Wen, Wenming Cao, Student Member, IEEE, C. L. Philip Chen , Fellow, IEEE, Hau-San Wong , and Jane You, “Hybrid Classifier Ensemble for Imbalanced Data*”, IEEE Transactions on Neural Networks and Learning Systems*, 31(4),pp.1387-1400, 2020.

[14] Pelin YJldJrJm, UlaGK.Birant, and Derya Birant, “EBOC: Ensemble-Based Ordinal Classification in Transportation”, *Journal of Advanced Transportation*, Volume 2019, pp.1-17, 2019.

[15] Tom M. Mitchell, “Machine Learning”, McGraw-Hill, New York, 1997.

[16] V. Vapnik, S. Golowich and A. Smola, “Support vector method for function approximation, regression estimation, and signal processing”, *Advances in neural information processing systems*, pp. 281-287, 1997.