Utkal University Journal of Computing & Communications,
Vol. 1, Issue. 2 (2023) *e-ISSN:2584-203X*

A Novel Ensemble Method for Imbalanced Data Classification

M. Govindarajan[1],
[1]Associate Professor, Department of Computer Science and Engineering,
Annamalai University,
Annamalai Nagar – 608002,
Tamil Nadu, India.
govind_aucse@yahoo.com

Abstract: Classification algorithms have shown exceptional prediction results in the supervised learning area. These classification algorithms are not always efficient when it comes to real-life datasets due to class distributions. As a result, datasets for real-life applications are generally imbalanced. Numerous methods have been developed to treat imbalanced datasets, which can be divided into three categories: (1) data resampling (2) algorithm modification and (3) ensemble methods. Among those categories, ensemble methods are the important area that proves to improve the classification performance. Ensemble learning combines several base models, where a traditional algorithm is used to learn each of them. It aggregates the outputs from a set of different classifiers to correctly classify new data points. Some popular ensemble learning methods include Bagging, Boosting and Adaboost. Bagging is an inherent parallel ensemble learning technology whose components can be running at the same time, and uses majority voting or weighted majority voting to aggregate results. It has provided considerable performance gains over a single learner in many application domains. This paper proposed an ensemble methods using automobile data by fusing classifiers such as RBF and SVM with bagging and their performances are analyzed in terms of accuracy. A wide range of comparative experiments are conducted for standard dataset of automobile. The proposed bagged ensemble methods provide significant improvement of accuracy compared to individual classifiers and previous works on standard dataset of automobile are exhibited.

Keywords: Accuracy, Bagging, Ensemble, Radial Basis Function, Support Vector Machine

1. Introduction

With the development of information technology and industry applications, the volume of data is increasing rapidly. It is a popular trend that adopting machine learning, artificial intelligence and deep learning to get latent information from data for providing users with more smart service. Traditional classification algorithms take the assumption that data has a good distribution; however, it is common that training data are imbalanced over classes, which leads to the bias of learning algorithms. The research on imbalanced classification has recently drawn much attention. Over the past two decades, many learning algorithms for addressing the imbalanced classification problem have been proposed. The methods mainly include sampling, cost-sensitive learning, threshold-moving, one-class learning and

✉ govind_aucse@yahoo.com

ensemble learning or multiple classifiers system. Ensemble learning (Roshan et al., 2020) aims to utilize the other class imbalance learning methods, e.g., sampling, cost-sensitive-cost learning or threshold-moving, to combine with the emerging ensemble learning paradigms, e.g., bagging or boosting, for classifying imbalanced data. In comparison to those single models, ensemble learning is expected to greatly improve the classification performance, especially the generalization ability, on class imbalanced data. In view of the merits of ensemble learning, it has been widely studied and adopted in the context of imbalanced data classification. The main contribution of this paper is to apply homogeneous ensemble classifiers using bagging 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.

2. Related Works

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

Yi Liu et al., (2019) proposed an ensemble classification method that combines evolutionary under-sampling and feature selection. The Bootstrap method is employed in original data to generate many sample subsets. V-statistic is developed to measure the distribution of imbalanced data, and it is also taken as the optimization objective of the genetic algorithm for the under-sampling sample subsets. Moreover, F1 and Gmean indicators are taken as two optimization objectives and employ the multiobjective ant colony optimization algorithm for feature selection of resampled data to construct an ensemble system. The experimental results show that our proposed system has a better classification performance compared with other algorithms, especially for the high-dimensional imbalanced data.

Goksu Tuysuzoglu et al., (2020) proposed a novel modified version of bagging, named enhanced Bagging (eBagging), which uses a new mechanism (error-based bootstrapping) when constructing training sets in order to cope with this problem. In the experimental setting, the proposed eBagging technique was tested on 33 well-known benchmark datasets and compared with both bagging, random forest and boosting techniques using well-known classification algorithms: Support Vector Machines (SVM), decision trees (C4.5), k-Nearest Neighbour (kNN) and Naive Bayes (NB). The results show that eBagging outperforms its counterparts by classifying the data points more accurately while reducing the training error.

Moussa Diallo et al., (2021) proposed a hybrid method combining the pre-processing techniques and those of ensemble learning. The original training set is under sampled by evaluating the samples by stochastic measurement (SM) and then training these samples selected by Multilayer Perceptron to return a balanced training set. The MLPUS (Multilayer perceptron under sampling) balanced training set is aggregated using the bagging ensemble method. The MLPUS (Multilayer perceptron undersampling) balanced training set is aggregated using the bagging ensemble method. This method is also compared with six other existing methods in the literature, such as the MLP classifier on the original imbalance dataset, MLPUS, UnderBagging (combining random undersampling and bagging), RUSBoost, SMOTEBagging (Synthetic Minority Oversampling Technique and bagging), SMOTEBoost. The results show that our method is competitive compared to other methods.

Xie et al., (2022) proposed a hierarchical ensemble method for improved imbalance classification. Specifically, the

first level ensemble based on bootstrap sampling with replacement is performed to create an ensemble. Then, the second-level ensemble is generated based on two different weighting strategies, where the strategy having better performance is selected for the subsequent analysis. Next, the third-level ensemble is obtained via the combination of two methods for obtaining mean and covariance of multivariate Gaussian distribution, where the oversampling is then realized via the fitted multivariate Gaussian distribution. Here, different subsets are created by (1) the cluster that the current instance belongs to, and (2) the current instance and its k nearest minority neighbors. Furthermore, Euclidean distance-based sample optimization is developed for improved imbalance classification. Finally, late fusion based on majority voting is utilized to obtain final predictions.

Md. Siraj-Ud-Doulah et al., (2023) used twelve machine learning algorithms are considered: NB, LDA, LR, ANN, SVM, K-NN, HT, DT, C4.5, CART, RF, BB and evaluated the performance of the machine learning algorithms on both binary and multiple classification problems using a variety of performance metrics: accuracy, kappa statistic, precision, recall, specificity, F-measure, MAE, RMSE and MCC. It is found that RF algorithm proved to have the best performance in three out of seven datasets. But the other four algorithms: NN, NB, BB and LR also performed well.

The performance of the proposed bagged RBF, bagged SVM classifiers are examined in comparison with individual classifiers and previous works on standard dataset of automobile are exhibited.

3. Methodology

3.1. Pre-processing

In the process of database preprocessing, transformation and cleaning are executed. Cleaning indicates removal of the unnecessary tags and fill up the missed values in the datasets. Transformation indicates translating task of complete dataset into the preferred form (it is conversion task of numeric values into the character data type).

3.2. Existing Classification Methods

3.2.1. RBF

To alter the distance from a certain place, RBF utilizes RBF for activation. The system control, time-series prediction, and classification are used by RBF for functional approximation. RBF was utilized for classifying record in non-linear mode and compared entering record with trained dataset. The weighted linear superposition is the manufacture of the RBF NN. In the RBF prototype, the Gaussian-basis procedure is the frequently used basis function. (D.S.Broomhead et al., 1988).

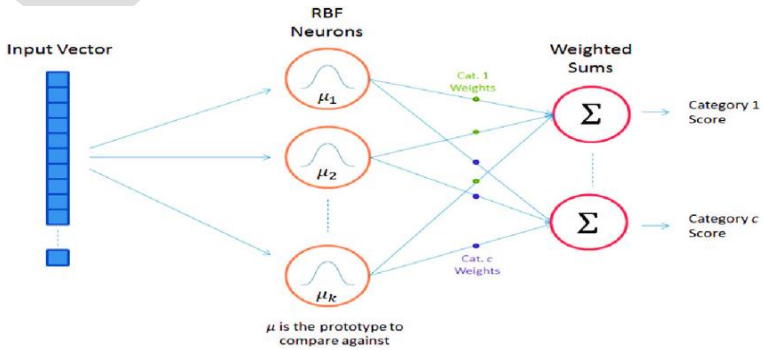

Fig. 1. RBF Network architecture

3.2.2. SVM

This is developed by J. Platt (1998) and broadly used to train SVM. SMO is a simple solution of solving a Quadratic Programming (QP) problematic issue that rises in the exercise process of SVM. SMO splits the huge QP issue into a sequence of minute sub-issues. Such minor sub-issues are methodically resolved, stopping the implementation of time-taking arithmetical QP maximization as an inner iteration. It was the rapid process for sparse datasets and linear SVM and comparatively faster than the chunking method. The retention looked-for SMO is linear in the exercise data scope, allows SMO to deal with huge exercise sets. It scales amongst quadratic and linear in the exercise set scope for various trial issues.

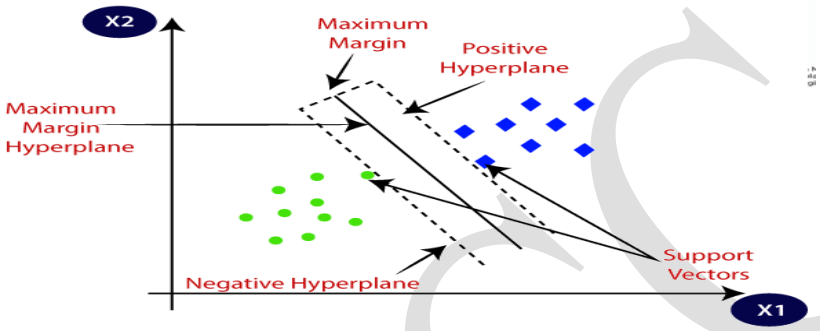

Fig. 2. SVM Architecture

3.3. Homogeneous Ensemble Classifiers (HMEC) Using Bagging

3.3.1. Proposed Bagged RBF and SVM Classifiers

Given a set D, of d tuples, bagging (Breiman, L. 1996) works as follows. For iteration i (i =1, 2,…..k), a training set, Di, of d tuples is sampled with replacement from the original set of tuples, D. The bootstrap sample, Di, 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 Di. A classifier model, Mi, is learned for each training set, Di. To classify an unknown tuple, X, each classifier, Mi, 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) End for

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;

Bagging performs well with the unstable algorithms (radial basis function, support vector machine) compared to stable learning algorithm

4. 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.

Table 1. Properties of Automobile Datasets

Datasets	Instances	Attributes
Vehicle	846	19

5. Experimental Results And Discussion

The proposed ensembles accuracy is evaluated to investigate the accomplishment of the homogeneous prototypes.

Table 2. Experimental Results for Homogeneous Ensemble Classification for Vehicle Dataset

Techniques	Accuracy Claimed
RBF	66.66%
SVM	74.34%
Homogeneous Group Classification	
Proposed Bagged RBF	**78.36%**
Dayvid V.R. Oliveira et al., 2017	71.62%
Wei Feng et al., 2018	71.20%
Shou Feng et al., 2020	69.60%
Moussa Diallo et al., 2021	75.37%
Xie et al., 2022	74.78%
Proposed Bagged SVM	**76.83%**
Bhavesh Patankar et al., 2015	75.17%
Samuel Giatti da Silva Filho et al., 2017	72.00%
Xiaobo Liu et al., 2018	75.62%
Goksu Tuysuzoglu et al., 2020	75.30%
Artittayapron Rojarath et al., 2021	73.23%
Xie et al., 2022	75.11%

In this research work, new ensemble classification method is proposed using bagging classifier in conjunction with radial basis function classifier and Support Vector Machine as the base learners 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 (Kohavi, R, 1995) technique is applied to the base classifiers and evaluated classification

accuracy. Bagging is performed with radial basis function classifier and Support Vector Machine to obtain a very good classification performance. Table 2 shows classification performance for auto mobile dataset using existing and proposed bagged radial basis function neural network and Support Vector Machine. The analysis of results shows that the proposed bagged radial basis function and proposed bagged Support Vector Machine are shown to be superior to individual approaches for auto mobile dataset in terms of classification accuracy. The $\chi 2$ 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 $\chi 2$ significance table, it is found that this value is significant with a degree of freedom of 1. In general, the result of $\chi 2$ statistic analysis shows that the proposed classifier is significant at $p < 0.05$ than the existing classifier. The results indicate that higher accuracy is achieved with the proposed bagged RBF and SVM in comparison to the base classifiers and prior work on standard dataset of automobile data set as given in Table 1.

6. Conclusıon

In this research, a new method of associating the categorization prototypes concerning homogeneous groups with bagging are employed by implementing vehicle information and the categorizer accomplishment is described with accuracy. Also, the proposed ensembles incorporate aspects of subsequent single categorizers.

The below comments are exposed from the outcomes.

- ❖ Amongst the standalone categorized employed, SVM describes crucially greater accomplishment in significant feature of accuracy.
- ❖ The bagged prototypes have been identified to accomplish outstanding improvement of categorization accuracy when related to the associating discrete categorizers.
- ❖ The fusion prototypes demonstrate crucially huge accuracy outcomes than collective prototypes on automobile database.

Developing and employing greatly literal categorizers precisely for the automobile database will be the upcoming study.

Acknowledgement
The author recognizes Annamalai University authorities for their esteemed assistance complete this study.

References

1. Artittayapron Rojarath & Wararat Songpan. (2021). Cost-sensitive probability for weighted voting in an ensemble model for multi-class classification problems. *Applied Intelligence*, 51, 4908–4932. https://doi.org/10.1007/s10489-020-02106-3
2. Bhavesh Patankar & Vijay Chavda. (2015). Improving Classification Accuracy through Ensemble Technique in Data Mining. *International Journal of Scientific Research in Science, Engineering and Technology*, 1(6), 193-197.
3. Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140. https://doi.org/10.1007/BF00058655
4. Broomhead, D.S. & Lowe, D. (1988). Radial basis functions, multi-variable functional interpolation and adaptive networks. *Complex Syst.*, 2, 321–355.

5. Dayvid V.R. Oliveira, George D.C. Cavalcanti, & Robert Sabourin. (2017). Online pruning of base classifiers for Dynamic Ensemble Selection. *Pattern Recognition*, 72, 44-58. https://doi.org/10.1016/j.patcog.2017.06.030
6. Goksu Tuysuzoglu & Derya Birant. (2020). Enhanced Bagging (eBagging): A Novel Approach for Ensemble Learning. *The International Arab Journal of Information Technology*, 17(4), 515-528. https://doi.org/10.34028/iajit/17/4/10
7. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of International Joint Conference on Artificial Intelligence*. 1137– 1143. https://doi.org/10.5555/1643031.1643047
8. Md. Siraj-Ud-Doulah, Md. Nazmul Islam. (2023). Performance Evaluation of Machine Learning Algorithm in Various Datasets. *Journal of Artificial Intelligence, Machine Learning and Neural Network*, 03(02), 14-32. https://doi.org/10.55529/jaimlnn.32.14.32
9. Moussa Diallo, Shengwu Xiong, Eshete Derb Emiru, Awet Fesseha, Aminu Onimisi Abdulsalami & Mohamed Abd Elaziz. (2021). A Hybrid MultiLayer Perceptron Under-Sampling with Bagging Dealing with a Real-Life Imbalanced Rice Dataset. *Information*, 12, 291: 1-21. https://doi.org/10.3390/info12080291
10. Platt, J. (1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. *Technical Report; MSR-TR-98-14.* Microsoft Research: Redmond, WA, USA.
11. Roshan, S. E & Asadi, S. (2020). Improvement of bagging performance for classification of imbalanced datasets using evolutionary multi-objective optimization. *Eng. Appl. Artif. Intell.*, 87 (2020), 103319. https://doi.org/10.1016/j.engappai.2019.103319
12. Samuel Giatti da Silva Filho, Roberto Zanetti Freire, & Leandro dos Santos Coelho. (2017). Feature Extraction for On-Road Vehicle Detection Based on Support Vector Machine. *proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. Bruges (Belgium), 65-70.
13. Shou Feng, Chunhui Zhao, & Ping Fu. (2020). A cluster-based hybrid sampling approach for imbalanced data classification. *Rev. Sci. Instrum*, 91: 1-8. https://doi.org/10.1063/5.0008935
14. Wei Feng, Wenjiang Huang & Jinchang Ren. (2018). Class Imbalance Ensemble Learning Based on the Margin Theory. *Applied Sciences*, *8*(5), 815-843. https://doi.org/10.3390/app8050815
15. Xiaobo Liu, Zhentao Liu, Guangjun Wang, Zhihua Cai, & Harry Zhang. (2018). Ensemble Transfer Learning Algorithm. *IEEE Access: Advanced Data Analytics for Large-scale Complex Data Environments*, 6, 2389 – 2396. https://doi.org/10.1109/ACCESS.2017.2782884
16. Xie, J., Zhu, M., & Hu, K. (2022). Hierarchical Ensemble Based Imbalance Classification, Computational Science – ICCS 2022. *Lecture Notes in Computer Science*, 13350, 192–204. https://doi.org/10.1007/978-3-031-08751-6_14
17. Yi Liu, Yanzhen Wang, Xiaoguang Ren, Hao Zhou & Xingchun Diao. (2019). A Classification Method Based on Feature Selection for Imbalanced Data. *IEEE Access*, 7, 81794 - 81807. https://doi.org/10.1109/ACCESS.2019.2923846