Ensemble Learning Technique to Improve Classification Accuracy for Credit Data

Jismy Joseph[1], Dr.G. Kesavaraj[2]
PhD Research Scholar[1], Professor and Head[2]
Department of Computer Science, Vivekanandha College of Arts and
Science for Women (Autonomous), Elayampalayam, Thiruchengode, Tamil Nadu, India

ABSTRACT

Now a day's, Ensemble methods are some of the most influential strategies in data mining and machine learning. It combines multiple learning algorithms into one, to obtain a more accurate predictive result. Credit risk analysis is one of the serious tasks in financial sector. By using ensemble methods the credit data can be classified more perfectly than by using a basic model. This paper represents a comparative study of different classifier on credit data set when the ensemble learning method 'Bagging' is used. This study observed that Bagging method can improve the accuracy of the basic classifier.

Keywords– Ensemble learning, Classification, Bagging, Machine Learning.

INTRODUCTION

Ensemble methods are powerful machine learning techniques that combine multiple basic classifiers and improve the accuracy of the predictive model. In recent years, various ensemble methods have been recommended such as boosting, bagging, voting etc, which have proved to be very valuable for machine learning and data mining [1], [2]. Bagging is used to reduce the variance of the decision tree. In bagging , a sample of a training data set which contains N observation and M features is taken randomly and the best split is used to split the node. This new training set is known as 'Bootstrap replicate'. Whereas boosting is used to create a set of predictors. Boosting creates a training set of learners, sequentially and combining them for prediction.

LITERATURE REVIEW

In [3] the study indicates that the bagging ensemble method can substantially improve individual base learners such as decision tree, multilayer perception, and k-nearest neighbors [3]. After applying ensemble method, the performance of SVM does not change. The results show that k-nearest neighbor is more appropriatefor large unbalanced datasets in credit scoring.

In [4] ,NP Singh, concluded that the data small or big should be subjected to many algorithms and their combinations using hybrid or esemble, produced more reliable output.

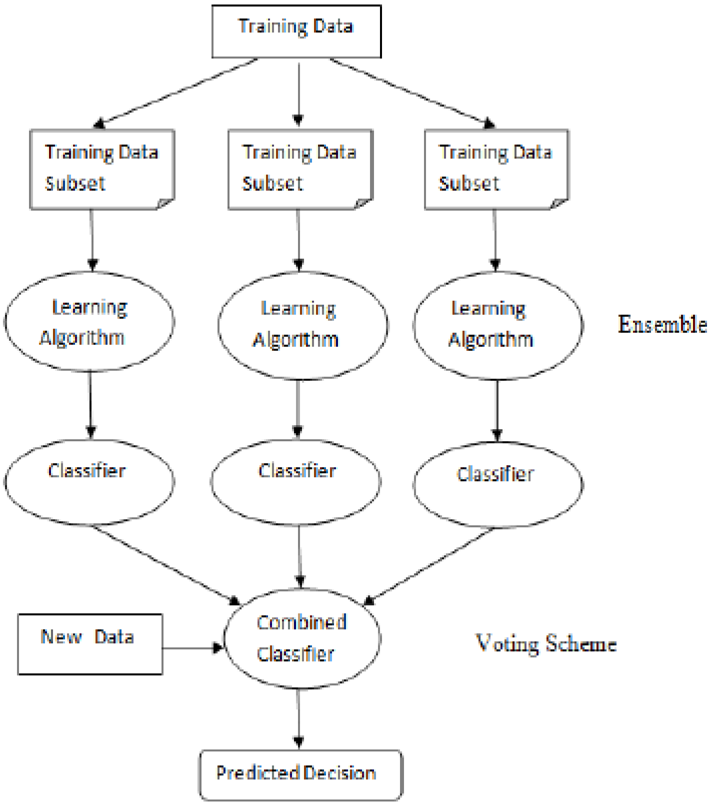
In [5], they compared base classifiers in ensemble methods for credit scoring and suggested that ensemble methodsprovides more suitable result for credit scoring..

In [6], The research compared the predictive accuracy of ensemble of base classifiers using techniques of bagging, boosting, and random forest in the prediction of default of credit card clients and suggested that Boosting ensemble technique is found to have the best accuracy of prediction.

In [7], their aim is to conduct empirical analysis on publically available bank loan dataset to study banking loan default using decision tree as the base learner and comparing it with ensemble tree learning techniques such as bagging, boosting, and random forests. The results indicate that ensemble model works better than the individual models.

ENSEMBLE METHOD - BAGGING

Bagging is also known as Bootstrap aggregation It is a machine learning algorithm used to improve the accuracy of the classification algorithms. It is mainly used in decision tree approaches. Bagging creates n new training set from a standard training data set of size m. The Bagging procedure is shown in Fig.1



(Fig.1 Bagging procedure)

DATA SET AND CLASSIFIERS

Three sets of credit data from UCI repository have been used for comparing five algorithms to find credit risk. The first data set is an Australian credit data set. This data set consists of 15 attributes and 690 instances. The second set is a Japanese credit data which has 16 attributes and 690 instances. The third one is a German credit Data set with 21 attributes and 1000 instances. In this comparative study fivebasicclassification algorithms and bagging method are used. The tool Weka is used to compare the accuracy of these basic classifiers with bagging. The classification algorithms used in the study are:

Random Forest

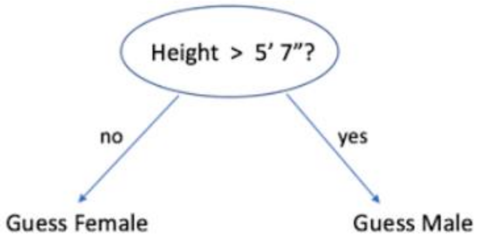
Random Forest is a supervised machine learning algorithm used for classification and regression. These classifiers handle the missing values and can model the **categorical values**. It creates many decision trees and merges them together to form an accurate prediction. In this method the parameters are used to increase the predictive power and speed of the model.

REPTree

REP Tree is a regression based classifier, it generates multiple trees in different iteration and selectsthe best one from these and is considered as the representative one.

Decision Stump

Decision stump is decision tree based machine learning model. This decision tree has only one root node and immediately connected to its leaves (terminal node). It makes a prediction based on the value a single input feature**.**



(Fig.2 - Processing of Decision Stump)

Random Tree

Random tree is a powerful method for image classification, mainly used for classification and regression problems. Random tree classifier classifies the input vector with every tree in the forest produce the class label that has maximum vote.

RESULT AND OBSERVATION

The text data is converted into ARFF format first, then data preprocessing is performed to create quality data. After that the classification is performed by using single basic classifier like REFTree,J48, Randomforest, Random tree and DecisionStump. Next the bagging classifier is selected and the single basic classifier(REFTree, J48, Randomforest, RandomTree and DecisionStump) are applied with bagging.

Table 1 shows that the accuracy of algorithm with bagging is greater than the the accuracy of the single classifier. The algorithm Random forest with bagging has highest accurate rate in all the data set and the RandomTree with Bagging shows the lowest accuracy in all the cases.

Table1 :- classification accuracy of different classifier with and without bagging

Classier	Australian Credit Data-Accuracy (%)	German Credit Data-Accuracy (%)	Japanese Credit Data-Accuracy (%)
J48	85.2174	70.7	86.087
J48 with Bagging	86.96%	74.70%	86.8116
Random Forest	86.9565	76.8	86.6667
Random Forest with Bagging	87.971	76.9	87.1014
REPTree	84.7826	71.8	85.6522
REPTree with Bagging	86.38%	74.70%	85.6522
RandomTree	79.86%	66.1	77.5362
RandomTree with Bagging	85.65%	74.9	84.6377
DecisionStump	85.51%	70	85.5072
DecisionStump with Bagging	85.51%	70	85.65

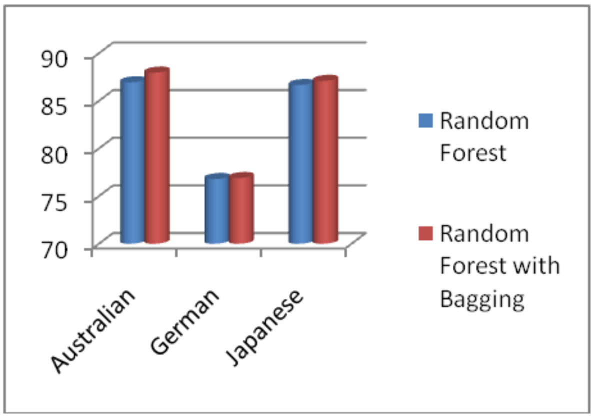


Fig3.(a)

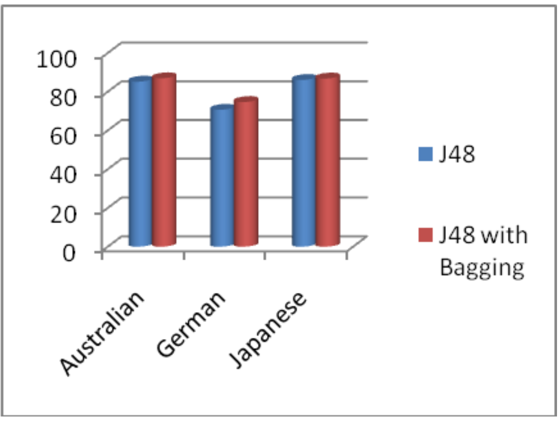


Fig3.(b)

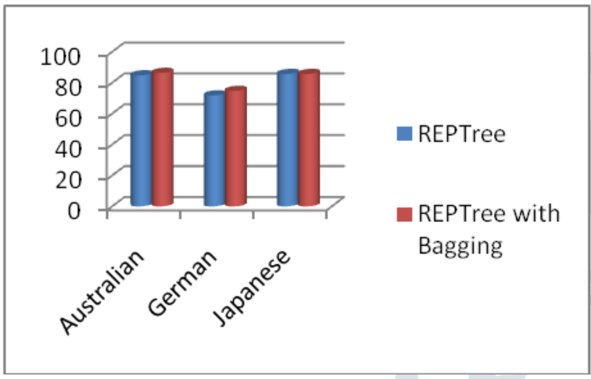


Fig3.(c)

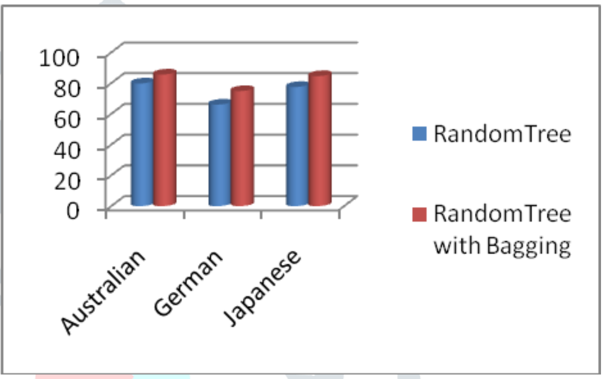


Fig3.(d)

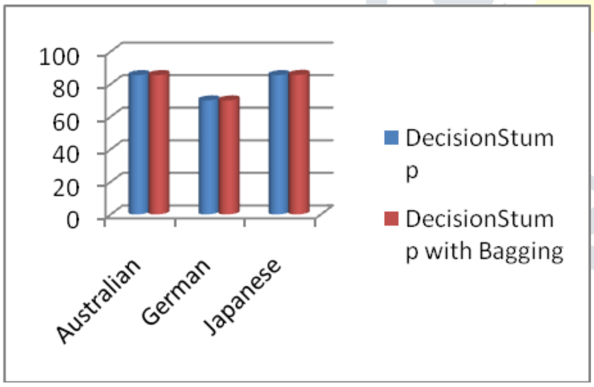


Fig3.(e)

(Fig 3. Comparative analysis of Random forest , J48, REF tree Random Tree and decision stump with Bagging)

CONCLUSION

This paper focuses on trying to find the effect of bagging when calculating the accuracy of different algorithms. This study observed that Random Forest algorithm with bagging obtained highest acuracy in all the data sets. It produced 87.9% of accuracy in Australian data set , 77% in german data set and 86% in japanese data set.

REFERENCES

1. Hatami, N. (n.d.). Thinned ECOC ensemble based on sequential code shrinking. expert system with applications .
2. Abellán, J., & Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring.Expert Systems with Applications, 73, 1-10.
3. Luo, C. (2018, June). www.researchgate.net/scientific-contributions/2143655443_Cuicui_Luo. Retrieved from www.researchgate.net: https://www.researchgate.net/scientific-contributions/2143655443_Cuicui_Luo
4. NP Singh, N. G. (2018). Comaparative analysis of data mining models for classification for small data set. IEEE conference. Researchgate.
5. Abellán, J., & Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring.Expert Systems with Applications, 73, 1-10.
6. Singh, B. E. R., & Sivasankar, E. (2019). Enhancing Prediction Accuracy of Default of Credit Using Ensemble Techniques. In First International Conference on Artificial Intelligence and Cognitive Computing (pp. 427-436). Springer, Singapore.
7. Chopra, A., & Bhilare, P. (2018). Application of Ensemble Models in Credit Scoring Models. Business Perspectives and Research, 6(2), 129-141.