Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2013 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

6-18-2013

Double Ensemble Approaches to Predicting Firms' Credit Rating

Jungeun Kwon
Korea University, jugugbim@gmail.com

Keunho Choi
Korea University, keunho@korea.ac.kr

Yongmoo Suh
Korea University, ymsuh@korea.ac.kr

DOUBLE ENSEMBLE APPROACHES TO PREDICTING FIRMS' CREDIT RATING

Jungeun Kwon, Business School, Korea University, Seoul, Republic of Korea, jugugbim@gmail.com

Keunho Choi* Business School, Korea University, Seoul, Republic of Korea, keunho@korea.ac.kr

Yongmoo Suh, Business School, Korea University, Seoul, Republic of Korea, ymsuh@korea.ac.kr

Abstract

Several rating agencies such as Standard & Poor's (S&P), Moody's and Fitch Ratings have evaluated firms' credit rating. Since lots of fees are required by the agencies and sometimes the timely default risk of the firms is not reflected, it can be helpful for stakeholders if the credit ratings can be predicted before the agencies publish them. However, it is not easy to make an accurate prediction of credit rating since it covers a variety of range. Therefore, this study proposes two double ensemble approaches, 1) bagging-boosting and 2) boosting-bagging, to improve the prediction accuracy. To that end, we first conducted feature selection, using Chi-Square and Gain-Ratio attribute evaluators, with 3 classification algorithms (i.e., decision tree (DT), artificial neural network (ANN), and Naïve Bayesian (NB)) to select relevant features and a base classifier of ensemble models. And then, we integrated bagging and boosting methods by applying boosting method to bagging method (bagging-boosting), and bagging method to boosting method (boosting-bagging). Finally, we compared the prediction accuracy of our proposed model to benchmark models. The experimental results showed that our proposed models outperformed the benchmark models.

Keywords: Credit rating, Ensemble, Bagging, Boosting

* Corresponding author
Tel: +82 2 3290 1945
Fax: +82 2 922 7220
E-mail: keunho@korea.ac.kr

1 INTRODUCTION

Credit ratings are assessments of creditworthiness of issuers and involve a hierarchical ranking process by which credit is classified into different risk categories (Ong 2002). A firm's credit rating is evaluated by credit rating agencies such as Standard & Poor's (S&P)[1], Moody's, Fitch Ratings, A. M. Best, and Dun & Bradstreet. Since the credit rating agencies require amounts of fees for their services, and sometimes the credit ratings provided by them periodically do not reflect the timely default risk of the firm, it may be beneficial for stakeholders to be able to predict credit ratings before the agencies publish the ratings (Kim and Ahn 2012). In other words, it is meaningful for them to develop a prediction model for credit rating in order to make them less reliant on credit rating agencies.

However, since the rating scale has a wide range from AAA to D (i.e., 22 ratings in the case of S&P), it is not easy to make an accurate prediction of credit rating. Therefore, most of previous researches have used a small number of classes as values of a target variable in their prediction model to increase prediction accuracy. Nevertheless, it would be better if we could make a prediction model with high accuracy while addressing multiple classes, since the stakeholders can get more exact credit rating from the model. Since ensemble models have been used to increase the classification accuracy, we employed the ensemble methods, which would result in better accuracy especially when a target variable has multiple classes as in credit rating domain.

In spite of this importance of the ensemble models, only a few researches have adopted the ensemble models to predict firms' credit ratings. Therefore, this study utilized ensemble models to predict firms' credit rating as one of multiple classes. In addition, we further integrated the two ensemble models, bagging and boosting, to improve the prediction accuracy, which can be explained as follows. Since the former has a limitation in that it makes a decision using the majority voting strategy without considering the weights of the results from each of base classifiers, and the latter in that the classifiers built from the training dataset prepared in later steps may be over-emphasizing instances which may be a noise, it had better to integrate them so that their limitation can be mitigated (Freud and Schapire 1996; Opitz and Maclin 1999).

The objective of this paper is to suggest two double ensemble approaches, 1) bagging-boosting and 2) boosting-bagging, each of which applies boosting method to bagging method, and bagging method to boosting method, respectively. To that end, we first conducted feature selection, using Chi-Square and Gain-Ratio attribute evaluators, with 3 classification algorithms (i.e., decision tree [DT], artificial neural network [ANN], and Naïve Bayes [NB]) to select relevant features and a base classifier of ensemble models. And then, we compared the classification accuracy of proposed model to benchmark models (i.e., DT, bagging and boosting). All the details about these experiments are given later in Section 5.

The rest of this paper is organized as follows. In the next section, we review recent researches which applied data mining techniques to predicting credit rating. And then, Section 3 describes our proposed models which integrate two ensemble methods (i.e., bagging and boosting). In Section 4, data and experimental design are described. Section 5 explains the results of our experiments and compares them. In section 6, we conclude the paper.

2 LITERATURE REVIEW

A variety of methods have been applied to classifying credit rating. This section reviews recent researches which applied data mining techniques to predicting credit rating. Chen and Cheng (2012) proposed two hybrid models: 1) FA-RS and 2) MEPA-RS, using Factor Analysis (FA) or Minimize Entropy Principle Approach (MEPA) together with Rough Set theory (RS). Their dataset covered 1,950 samples including 420 large banks and 5 classes from the credit rating of Fitch. The accuracy of

[1] In the case of S&P, the rating scale is as follows, from best to worst: AAA, AA+, AA, AA-, A+, A, A-, BBB+, BBB, BBB-, BB+, BB, BB-, B+, B, B-, CCC+, CCC, CCC-, CC, C, and D.

FA-RS was 78.55% with 9 attributes and 79.29% with 6 attributes, and that of MEPA-RS was 82.14% with 16 attributes.

Kim and Ahn (2012) developed Ordinal Multi-class Support Vector Machine (OMSVM) which consists of two phases: 1) preparation which includes One-Against-The-Next and One-Against-Followers, and 2) interpretation which includes forward and backward. They applied this method to the instances with 14 variables of 1,295 companies from the manufacturing industry in Korea. The accuracy of OMSVM which classifies instances into one of four classes was 67.98%. Chen (2012) suggested a Cumulative Probability Distribution Approach (CPDA) and rough sets local-discretization cuts (RS) to partition selected condition attributes, and then adopted the rough sets (LEM2 algorithm) to generate a comprehensible set of decision rules. The dataset included 18 attributes from Asian banks and classified Fitch's credit rating into 5 classes. The accuracy was 81.75% with 18 attributes and 83.84%with 16 attributes.

Hájek (2011) applied various neural networks such as Feed-Forward NN (FFNN), Radial Basis Function NNs (RBFNNs), Probabilistic Neural Networks (PNNs), Cascade Correlation NNs (CCNNs), Group Method of Data Handling (GMDH) Polynomial NNs, and Support Vector Machines (SVMs). This study used 14 attributes from 169 US municipalities and classified Moody's credit rating into 9 classes or 4 classes. The highest classification accuracy, 98.8%, was obtained using PNN with 4 classes, and the next highest accuracy, 96.3%, using PNN with 9 classes. Cao et al. (2006) applied SVM with different sets of target classes such as one-against-all, one-against-one, and Directed Acyclic Graph SVM (DAGSVM) to the instances with 18 attributes to classify them into one of 6 classes from S&P's credit rating. The classification accuracies were 84.61%, 81.73% and 82.69% in DAGSVM, SVM with one-against-all, and SVM with one-against-one, respectively. Huang et al. (2004) applied SVM and back propagation neural network (BNN) to the samples with 21 attributes from the US (265 cases) and Taiwanese (74 cases) financial markets and classified bond ratings into one of 5 classes. The best classification accuracies of SVM and BNN were 80.38% and 80.75%, respectively.

Ensemble models have been widely used in a number of researches. However, to the best of our knowledge, only a few researchers have built ensemble models to classify the credit rating (Yeh et al. 2012; Ye et al. 2008). Yeh et al. (2012) proposed hybrid KMV model (developed by the KMV Corporation) which combine random forests (RF) and rough set theory (RST) to extract useful rules for credit rating. The dataset included instances of 2,470 Taiwanese high-technology companies with 22 attributes. The accuracy of the KMV model against 3 classes was 93.4%. Ye et al. (2008) applied bagged decision tree, multiclass SVM, and multiclass proximal SVM to 2,541 samples having 33 attributes. They classified Moody's issuer credit ratings into one of 19 classes. The best classification accuracies of bagged decision tree, multiclass SVM (MSVM), and proximal multiclass SVM (PMSVM) was 85.9%, 85.2%, and 84%, respectively. The fore-mentioned credit rating researches are summarized in Table 1.

References	Method	Samples	Attributes	Classes
Chen and Cheng (2012)	FA-RS, MEPA-RS	1,950	37	5
Kim and Ahn (2012)	OMSVM	1,295	14	4
Chen (2012)	CPDA, RS	1,327	18	5
Hájek (2011)	FFNN, RBFNN, PNN, CCNN, GMDH, SVM.	169	14	4, 9
Cao et al. (2006)	SVM (One-against-all), SVM (One-against-one), DAGSVM	239	18	6
Huang et al. (2004)	SVM, BNN	339	21	5

Yeh et al. (2012)	KMV	2,470	22	3
Ye et al. (2008)	bagged DT, MSVM, proximal MSVM	2,541	33	19

Table 1. Credit rating researches

3 PROPOSED DOUBLE ENSEMBLE APPROACHES

As mentioned in Section 2, ensemble models have been widely used in a number of researches to increase the prediction accuracy. However, to the best of our knowledge, only a few researches have addressed the ensemble models to predict credit ratings. Since ensemble models make a decision after combining the results of several base classifiers, the results from them are more reliable and accurate than those of a plain model which makes a decision solely based on a single classifier. Especially when a target variable has multiple classes as in our study, we thought, it is an appropriate approach to build ensemble models to predict firms' credit ratings.

A bagging model, one of ensemble models, combines multiple base classifiers built from multiple training datasets each of which is obtained by sampling with replacement from an original single training dataset. And then, it makes a final decision using the majority voting of the multiple base classifiers. This bagging method, however, has a limitation in that it makes a decision using the majority voting strategy without considering the weights of the results from base classifiers. For example, when combination of three base classifiers, two with lower accuracy and the other with higher accuracy, predicts a new sample using the majority voting strategy, the prediction may give an inaccurate result.

A boosting model also combines multiple base classifiers from multiple training datasets. However, unlike the bagging method, each base classifier is built by focusing on the previous one's errors. That is, instances which are misclassified by previous base classifier are more likely to be selected in a training dataset when building the next base classifier. This boosting method makes a final decision using the weighted average of the multiple base classifiers. Although this aspect of the boosting method can reduce the number of misclassification, it can also incur over-fitting problem. Since the classifiers built from the training dataset prepared in later steps may be over-emphasizing instances which may be a noise, the boosting model can result in poor accuracy (Freud and Schapire 1996; Opitz and Maclin 1999).

With the expectation that the above limitations (or the problems) of the bagging and boosting methods can be mitigated by integrating them, we proposed two double ensemble approaches, 1) bagging-boosting and 2) boosting-bagging, each of which applies boosting method to bagging method and bagging method to boosting method, respectively (see Figs. 1 and 2, respectively).

As shown in Fig. 1 which explains bagging-boosting approach, we first make sub-training datasets as is done in the bagging method and then we make sub-sub-training datasets from each of sub-training datasets as is done in the boosting method. The final decision (or prediction) is made using the majority voting (as in bagging) of the weighted averages (as in boosting) of the predictions by base classifiers with sub-sub-training datasets.

As shown in Fig. 2 which explains boosting-bagging approach, we first make sub-training datasets as is done in the boosting method and then we make sub-sub-training datasets from each sub-training dataset as is done in the bagging method. The final decision (or prediction) is made using the weighted average (as in boosting) of the majority votes (as in bagging) by base classifiers with sub-sub-training datasets.

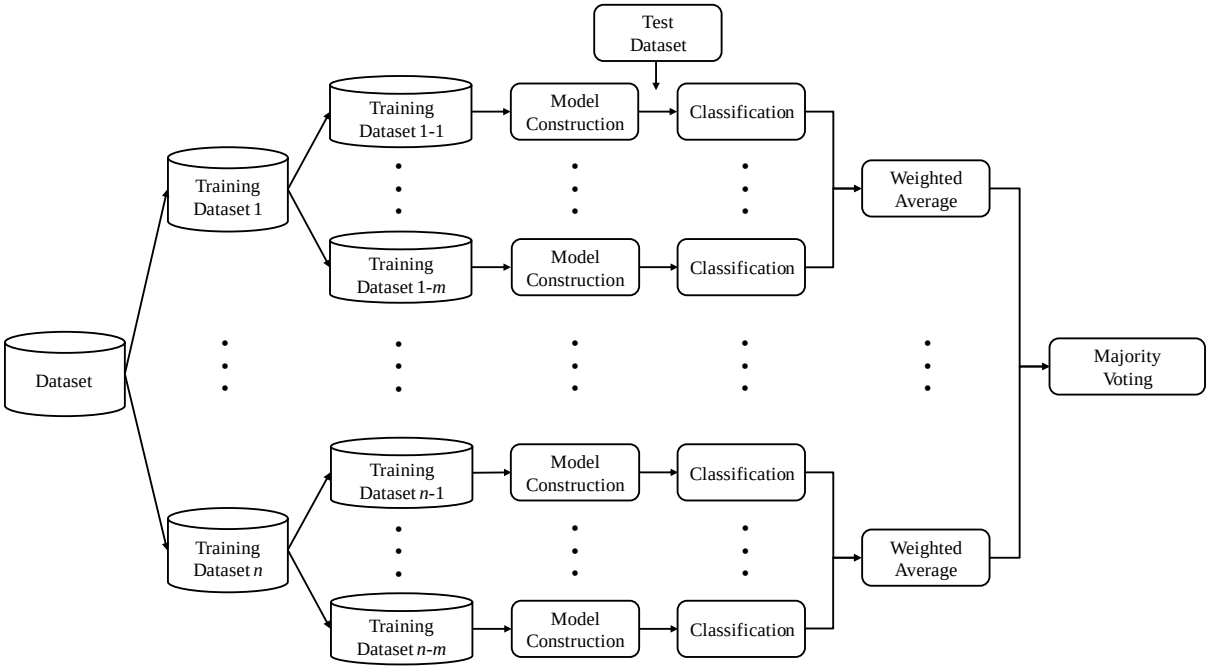


Figure 1. Bagging-boosting approach

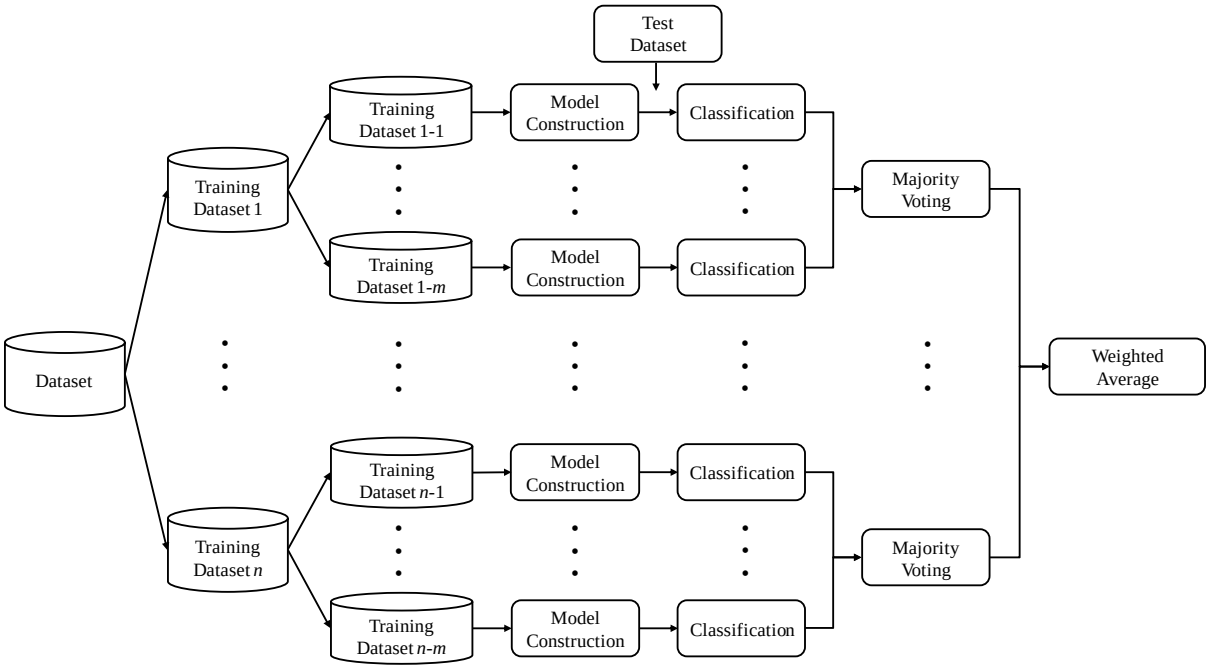


Figure 2. Boosting-bagging approach

4 EXPERIMETNS

4.1 Data

Our experimental data consists of two datasets both of which are obtained from Wharton Research Data Services (WRDS) which is the leading data research platform and provides business intelligence tool for over 30,000 corporate, academic, government and nonprofit clients across 30 countries. It covers the period from 2002 to 2012 and includes 1,480 companies from various countries.

The first dataset contains the rating information of companies evaluated by Standard & Poor's (S&P). S&P evaluated credit ratings of those firms as one of 22 classes. We reduced the number of classes from 22 to 16, since the number of sample belonging to the 6 classes (i.e., CCC+, CCC, CCC-, CC, C,

and D) are very small. The second dataset includes financial information of the same companies as those which appear in the first dataset of credit rating information. Among the financial information, we selected 43 financial variables that were known to have an effect on credit ratings in the previous related studies. Then, we joined the two files of financial information and credit ratings into one and eliminated records and columns with many missing values. Consequently, we obtained 21,321 samples and 31 financial variables. The 31 financial variables we obtained are listed in Table 2, and they will be used as inputs when building a model which predicts the credit rating of companies.

Variables	Definition
ACOQ	Current Assets - Other – Total
AOQ	Assets - Other – Total
APQ	Account Payable/Creditors – Trade
ATQ	Assets – Total
CEQQ	Common/Ordinary Equity – Total
CHEQ	Cash and Short-Term Investments
COGSQ	Cost of Goods Sold
DLCQ	Debt in Current Liabilities
DLTTQ	Long-Term Debt - Total
EPSF12	Earnings Per Share 12 Months Moving (Diluted) - Excluding Extraordinary Items
EPSFXQ	Earnings Per Share Quarterly (Diluted) - Excluding Extraordinary Items
EPSPXQ	Earnings Per Share Quarterly (Basic) - Excluding Extraordinary Items
EPSX12	Earnings Per Share 12 Months Moving (Basic) - Excluding Extraordinary
IBQ	Income Before Extraordinary Items
INTANQ	Intangible Assets - Total
INVTQ	Inventories - Total
LCOQ	Current Liabilities - Other - Total
LOQ	Liabilities - Other
LTQ	Liabilities - Total
NIQ	Net Income (Loss)
NOPIQ	Non-Operating Income (Expense) - Total
OIADPQ	Operating Income After Depreciation - Quarterly
PIQ	Pretax Income
PPENTQ	Property Plant and Equipment - Total (Net)
PSTKQ	Preferred/Preference Stock (Capital) - Total
RECTQ	Receivables - Total
SALEQ	Sales/Turnover (Net)
SEQQ	Stockholders Equity
SPIQ	Special Items
TXTQ	Income Taxes - Total
XINTQ	Interest and Related Expense- Total

Table 2. Financial features for credit rating

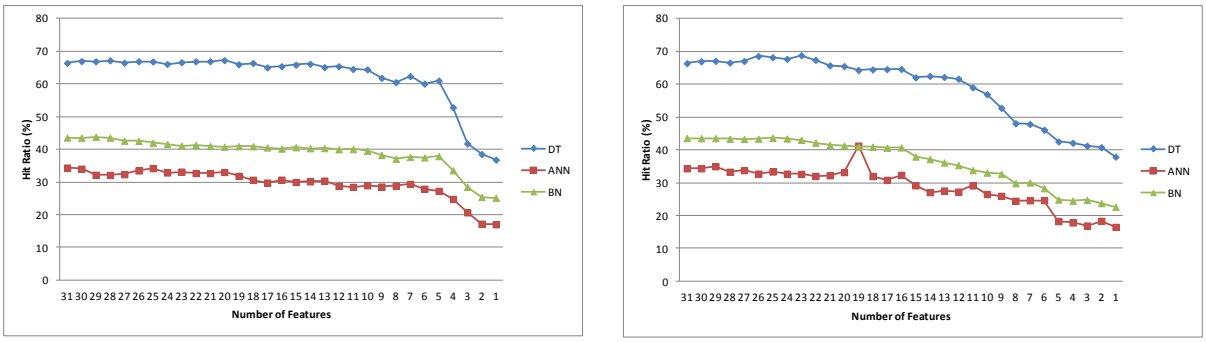
4.2 Experimental design

In this study, we conducted experiments with the same size of instances (i.e., 1,000) for each of 16 classes using over-sampling and under-sampling to resolve the problem of data imbalance. After

balancing, as a consequence, we obtained 16,000 instances. To build classification models, the dataset is randomly split into two sub-datasets, 70% for training and 30% for validation.

In our experiment, we used WEKA ver. 3.6.6 as a data mining tool, which is open source software and has been used widely. Prior to the construction of classification models, we conducted feature selection. We evaluated 31 input variables using Chi Square and Gain Ratio attribute evaluators based on ranker search method, to rank influential variables on the target variable. With the results of evaluation, we adopted stepwise backward elimination method which is used to get the most suitable subset of attributes by building classification models, first using the whole set of attributes and then removing the least influential variable one by one. We built classification models using decision tree (DT), artificial neural network (ANN), and Naïve Bayes (NB) algorithms. The parameters of all classification algorithms were set to the default values in WEKA. The performance comparison among these algorithms in our experiment was made based on the hit-ratio.

Figure 3 shows the hit ratios of the three classification models, changing as the number of features decreases in stepwise backward elimination. How influential each financial variable is on the credit rating is evaluated using Chi Square and Gain Ratio methods. As shown in the figure, we can see that DT shows the best performance among the three models, regardless of the number of features. More specifically, the highest classification accuracy of DT model was acquired when we used 20 attributes for Chi Square (67.21%) and 23 attributes for Gain Ratio (68.79%). In the case of Chi Square, ANN and NB show their highest classification accuracy (34.35% and 43.79%, respectively) with 31 and 29 attributes, respectively. In the case of Gain Ratio, ANN and NB show their highest classification accuracy (34.98% and 43.63%, respectively) with 29 and 25 attributes, respectively. Note that DT among the three classification algorithms shows the highest classification accuracy even with the smallest number of attributes, irrespective of the two feature evaluation methods. Therefore, we adopted DT as a base classifier of our proposed classification models: bagging-boosting and boosting-bagging approaches. The final attributes selected using Chi Square and Gain Ratio methods are listed in Table 3.



(a) Chi-Square (b) Gain-Ratio

Figure 3. Hit ratios of classification models as the number of features decreases

No.	Gain-Ratio	Chi-Square	No.	Gain-Ratio	Chi-Square
1	PIQ	CEQQ	13	LTQ	APQ
2	OIADPQ	SEQQ	14	APQ	SALEQ
3	NIQ	NIQ	15	ACOQ	DLTTQ
4	IBQ	INTANQ	16	PPENTQ	TXTQ
5	TXTQ	PPENTQ	17	CHEQ	LOQ
6	CEQQ	OIADPQ	18	LCOQ	INVTQ
7	PSTKQ	ATQ	19	COGSQ	ACOQ
8	SEQQ	IBQ	20	INVTQ	XINTQ

9	RECTQ	LTQ	21	LOQ	-
10	SALEQ	RECTQ	22	EPSX12	-
11	ATQ	AOQ	23	INTANQ	-
12	AOQ	PIQ	-	-	-

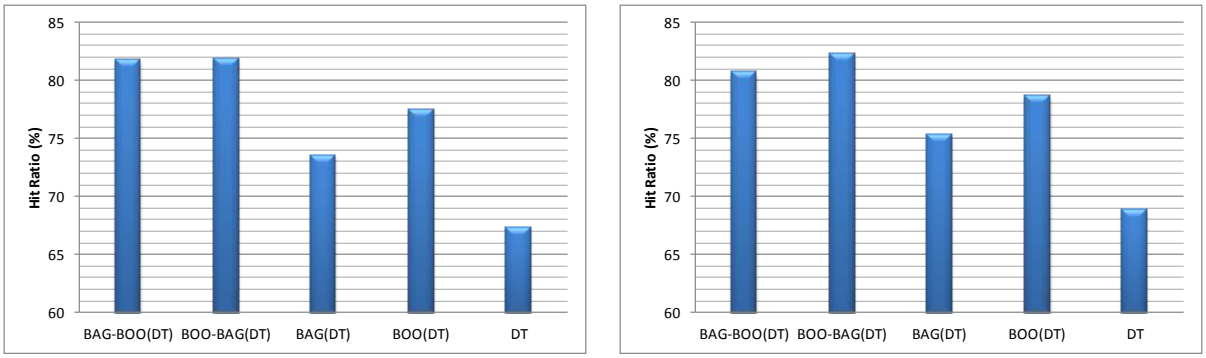
Table 3. Selected features using gain-ratio and chi-square algorithms, respectively

5 EXPERIMENTAL RESULTS

To verify the effectiveness of our proposed classification models, we used DT, bagging (DT) and boosting (DT) models as benchmark models. The parameters of the bagging method such as the size of each bag and the number of iterations were set to 90 and 9, respectively, in WEKA. For the boosting method, we used AdaBoost algorithm which has been used by many researches. And DT is used as a base classifier when building all ensemble models, proposed or benchmark.

Figure 4 shows the hit ratio of proposed classification models and benchmark models using the selected attributes (20 from Chi Square and 23 from Gain Ratio). As shown in this figure, we can see that both bagging-boosting (DT) and boosting-bagging (DT) methods proposed in this study outperformed the benchmark classification models, irrespective of the two feature evaluation methods. More specifically, Boosting-bagging (DT) yields the best performance (i.e., 82.27% for Chi Square, and 81.88% for Gain Ratio), which is a little bit better than either of Bagging-boosting (DT) (i.e., 81.71% for Chi Square and 80.73% for Gain ratio).

When Chi Square is used to evaluate the features, the classification accuracy of bagging (DT), boosting (DT), and DT were 73.52%, 77.44%, and 67.21, respectively, while when Gain Ratio is used, the classification accuracy of bagging (DT), boosting (DT), and DT were 75.33%, 78.58%, and 68.79, respectively.



(a) Chi-Square (b) Gain-Ratio

Figure 4. Hit ratio of proposed classification models and benchmark models

We calculated the improvement rate of classification models including the benchmark models to a base classifier, DT. The results are summarized in Table 4.

Base classifier	Feature Selection	**BAG-BOO (DT)**	**BOO-BAG (DT)**	BAG (DT)	BOO (DT)
DT	Chi-Square	21.57%	21.82%	9.39%	15.22%
	Gain-Ratio	17.35%	19.59%	9.51%	14.23%

Table 4. Improvement rate of classification models to a base classifier

6 CONCLUSIONS

In this paper, we proposed double ensemble models which integrate the bagging and the boosting methods in order to predict firms' credit ratings and improve the prediction accuracy. To verify the effectiveness of our prediction models, we conducted several experiments to compare the prediction accuracy of our proposed models with benchmark models. From the results, we can see that both bagging-boosting (DT) and boosting-bagging (DT) methods proposed in this study outperformed the benchmark models, regardless of the feature evaluation methods.

The proposed models achieved better prediction accuracy than benchmark models, and we expect that the proposed models can give benefits to stakeholders by predicting firms' credit rating more specifically and accurately, which is the contribution of this paper. Our study has a limitation in that our proposed approaches have high computational complexity. Nonetheless, we believe that it is worthwhile to apply the resulting models from our experiments to the credit rating domain.

REFERENCES

Cao, L., Guan, L. K., and Jingqing, Z. (2006). "Bond rating using support vector machine." Intelligent Data Analysis 10(3): 285-296.

Chen, Y.-S. (2011). "Classifying credit ratings for Asian banks using integrating feature selection and the CPDA-based rough sets approach." Knowledge-Based Systems 26(1): 259–270.

Chen, Y.-S. and C.-H. Cheng (2012). "Hybrid models based on rough set classifiers for setting credit rating decision rules in the global banking industry." Knowledge-Based Systems 39(1): 224-239.

Freund, Y. and R. E. Schapire (1996). "Experiments with a new boosting algorithm." In Proceedings of the thirteenth international conference on machine learning, Bari, Italy (pp. 148–156).

Hájek, P. (2011). "Municipal credit rating modelling by neural networks." Decision Support Systems 51(1): 108-118.

Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., and Wu, S. (2004). "Credit rating analysis with support vector machines and neural networks: a market comparative study." Decision Support Systems 37(4): 543-558.

Kim, K.-J. and H. Ahn (2012). "A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach." Computers & Operations Research 39(8): 1800-1811.

Maclin, R. and D. Opitz (2011). "Popular ensemble methods: An empirical study." Journal of Artificial Intelligence Research 11: 169-198.

M. Ong (2002) Credit Ratings: Methodologies, Rationale and Default Risk, Risk Books, British.

Ye, Y., Liu, S., and Li, J. (2008). "A Multiclass Machine Learning Approach to Credit Rating Prediction." Information Processing (ISIP), 2008 International Symposiums on, IEEE.

Yeh, C.-C., Lin, F., and Hsu, C.-Y. (2012). "A hybrid KMV model, random forests and rough set theory approach for credit rating." Knowledge-Based Systems 33(4): 166-172.