Multiple proportion case-basing driven CBRE and its application in the evaluation of possible failure of firms

Hui Li*[,1], Diego Andina[2], Jie Sun [1]

1. School of Economics and Management, Zhejiang Normal University, P.O. Box 62, 688 YingBinDaDao, Jinhua, Zhejiang 321004, PR China

2. Head of Group for Automation in Signal and Communications, Technical University of Madrid, ETSI Telecomunicación, Madrid 28040, Spain

Biography

Hui LI, received his BSc degree in Accounting, MS degree in Accounting, and a PhD degree in Management Science and Engineering from Harbin Institute of Technology, Harbin, Heilongjiang, China. The author's major field of study includes case-based reasoning, business forecasting, business computing, business data mining, and business intelligence, among others. He is an associate professor at Zhejiang Normal University, Jinhua, Zhejiang, China, and was a visiting scholar of College of Engineering, Ohio State University, Columbus, Ohio, United States. His researches were published in *Applied Soft Computing*, *Computers and Operations Research, European Journal of Operational Research, Expert Systems with Applications, Information & Management, Information Sciences, Journal of Forecasting, Knowledge-Based Systems,* among others. Dr. Li is a young researcher in the World Federation on Soft Computing, and a member of Associations for Information Systems. He serves as principal investigator of several national funded research projects, including: the National Natural Science Foundation of China and the Zhejiang Provincial Natural Science Foundation of China. He received the Outstanding Young Talent Award of Zhejiang Province in 2009.

Multiple proportion case-basing driven CBRE and its application in the evaluation of possible failure of firms

Abstract. Case-based reasoning (CBR) is a unique tool for the evaluation of possible failure of firms (EOPFOF) for its eases of interpretation and implementation. Ensemble computing, a variation of group decision in society, provides a potential means of improving predictive performance of CBR-based EOPFOF. This research aims to integrate bagging and proportion case-basing with CBR to generate a method of proportion bagging CBR for EOPFOF. Diverse multiple-case bases are firstly produced by multiple case-basing, in which a volume parameter is introduced to control the size of each case base. Then, the classic case retrieval algorithm is implemented to generate diverse member CBR predictors. Majority voting, the most frequently used mechanism in ensemble computing, is finally used to aggregate outputs of member CBR predictors in order to produce final prediction of the CBR ensemble. In an empirical experiment, we statistically validated the results of the CBR ensemble from multiple case bases by comparing them with those of multivariate discriminant analysis, logistic regression, classic CBR, the best member CBR predictor, and bagging CBR ensemble. The results from Chinese EOPFOF prior to 3 years ago indicate that the new CBR ensemble, which significantly improved CBR's predictive ability, outperformed all the comparative methods.

Keywords: Evaluation of possible failure of firms (EOPFOF); case-based reasoning ensemble; multiple case-basing; proportion bagging; nearest neighbor ensemble

1 Introduction

Case-based reasoning (CBR) is an easily interpretable and implementable methodology for problem-solving. The basic assumption of CBR is that similar cases have a similar outcome (Chang et al., 2006; Yankov, et al., 2006; Beddoe and Petrovic, 2006; Liu et al., 2008; Lee 2008; Zhuang et al., 2009; Castro et al, 2009; Ahn and Kim 2009). Case base and reasoning process are important in the use of solutions of similar cases to solve the target case. According to Aamodt and Plaza (1994), CBR consists of case retrieval, case reuse, case revision and case retaining. CBR is an effective method in evaluation of possible failure of firms (EOPFOF), which refers to identifying failed firms from non-failed ones by constructing approaches or models on financial or non-financial

information. EOPFOF is an effective tool in helping the people involved to make precise decisions in the current competitive environment. Thus, the issue of improving predictive performance of CBR in EOPFOF is critical. Case representation and case retrieval are two critical processes that determine output of CBR. The most commonly used algorithm for case retrieval is *k* nearest neighbor (kNN). In order to address the issue of making CBR more useful in EOPFOF, former researches have focused on researching into how to improve the predictive performance of CBR from the viewpoint of case retrieval (e.g., Li and Sun, 2008; Park and Han, 2002; Yip, 2004; Lin et al., 2009) and case representation (Li and Sun, 2010). It is of interest to investigate whether some alternative approaches are useful in helping CBR improve its performance.

The generalization abilities of ensemble models could be better than those of single predictors (Zhou and Yu, 2005). Bagging is one of the most easily implemented, widely used, classic and famous algorithms. This approach is used to generate various training datasets, with which member predictors are trained (Breiman, 1996), from an original dataset by bootstrap sampling with replacement. Lots of single predictive techniques, including: neural network and decision tree have been successfully aggregated with bagging to produce ensembles with a relatively higher performance. The so-called training datasets in ensemble refer to case bases in CBR. Thus, the use of multiple case bases generated by bagging is possible to produce diverse member CBR predictors, which is an alternative way of improving performance of CBR in EOPFOF. Skalak (1996), Bao and Ishii (2002) provided some evidence that ensembles of kNN generated better performance on some small datasets. An early integration on integration of the kNN algorithm with bagging indicated that bagging the kNN did not lead to increased accuracy on large datasets (Breiman, 1996; Alpaydin, 1997). The reason may be that the kNN algorithm is stable and errors made by individual kNN are highly correlated. For EOPFOF, both accuracy and the characteristic of ease of interpretation, explanation, and implementation are equally important. Thus, the issue that needs to be further

addressed is the construction of a CBR ensemble model that can provide an accurate prediction of business failure as well as the ease of interpretation, explanation, and implementation.

This research devotes an early investigation into the prediction of business failure with a CBR ensemble (CBRE) through the aggregation of CBR with bagging and proportional case-basing, whose method retains CBR's characteristic of ease of interpretation, explanation, and implementation when improving its predictive performance. It also makes bagging applicable to CBR with kNN at its heart by constructing diverse case bases. In order to produce diverse multiple case bases, we introduce a volume parameter to control the size of each case base. This parameter is integrated with the classic bagging algorithm to control volumes of multiple case bases, which is called proportion bagging. An empirical research was implemented to investigate how many random samples should be used to make up each case base of member CBR predictor for Chinese EOPFOF prior to 3 years ago. The empirical research was also conducted to investigate whether or not the proportion bagging CBR ensemble is effective for EOPFOF in comparison to multivariate discriminate analysis (MDA), logistic regression (logit), the classic CBR, the best member CBR, and bagging CBR. This research is organized as follows. Section 2 presents a short review of the CBR and kNN ensembles. Section 3 outlines the CBR ensemble from the view of how to construct diverse multiple case bases for member CBR predictors. Section 4 designs the empirical research to investigate the parameter range of proportion bagging and forecast Chinese business failure. Section 5 describes the results and gives some discussions. Section 6 presents conclusion and limitations.

2 Related work

In comparison with cautious focus on the CBR ensemble, numerous researches focusing on hybrid CBR and its applications. CBR is frequently combined with rule-based reasoning, fuzzy set, rough set, genetic algorithm and decision tree, among others, to solve real-world problems. For example, Chang et al., (2008a) developed a case-based evolutionary identification model for printed circuit board defect classification. The hybrid approach,

consisting of the referential process from CBR and the rule-based process from rule-based reasoning, attempts to take advantage of and overcome the drawbacks of each method. In order to find more suitable cases from case bases, Chang et al., (2008b) integrated fuzzy theories with CBR to produce a more flexible and accurate similarity model, and explored its potential application in sales forecasting in printed circuit board industries. Salamo & Lopez-Sanchez (2011) hybridized rough set approach with CBR for feature selection, where the rough set is supposed to help CBR manage imprecise, uncertain and redundant data. Ahn and Kim (2009) simultaneously optimized feature weighting and instance selection for CBR with genetic algorithm when solving the problem of bankruptcy prediction. The hybrid CBR is capable of improving its predictive performance by referencing more relevant cases. Fan et al., (2011) developed a hybrid model by combining CBR with a fuzzy decision tree to classify medical data. Empirical results indicate that the hybrid model produced the best performance in medical data classification.

Though the CBR ensemble is cautiously researched, some pioneer research provides useful knowledge. Up until now, three schemes have been successfully proposed in the CBR or kNN ensembles, which include: the generation of data subsets, the generation of feature subsets, and the generation of distance functions. The first scheme refers to the use of diverse training datasets to generate member predictors in CBR or kNN ensemble. Skalak (1996) and Bao and Ishii (2002) provided some evidence that the use of sample selection in generating member kNN predictors is supposed to improve the predictive performance of the ensemble in comparison with the single nearest neighbor predictors. The second scheme refers to generating individual kNN predictors with diverse feature subsets. The employment of diverse feature subsets is supposed to make the kNN unstable, since the CBR with kNN is sensitive to features (Pal and Shiu, 2004). Bay (1998) constructed a kNN ensemble by generating member kNN predictors on top of randomly generated subsets of original features and demonstrated performance improvement. Cunningham and Zenobi (2001) generated member kNN predictors from different

feature subsets, with the claim that this treatment helped the CBR ensemble produce a better predictive performance. Okun and Priisalu (2005) used multiple viewpoints with different features in ensembles of kNN algorithms, with the demonstration that the combination produced promising results. The third scheme refers to the generation of member predictors from the use of diverse distance functions. The commonly used distance function in kNN is Euclidean distance. This function is a specific form of Minkowsky distance. Thus, generation of diverse distance functions is supposed to provide an alternative for kNN ensemble. Bao et al. (2004) constructed a kNN ensemble by founding it on member predictors that have access to distance functions. Zhou and Yu (2005) adapted bagging to kNN algorithms by perturbing datasets and distance functions. A variation in using diverse distance functions in the CBR ensemble with kNN as the heart is to use diverse case retrieval techniques. Li and Sun (2009) used four different techniques to implement case retrieval and then constructed the CBR ensemble, with results indicating that the ensemble provided a more dominating performance. The use of the classic kNN algorithm in the ensemble, some variations of the algorithm are also useful. Altunçay (2007) combined the feature sub-spacing treatment with an evidential kNN algorithm in the construction of the kNN ensemble, with the result showing that an improved accuracy was obtained.

From the above review we find that there is both positive and negative evidence on the construction of the CBR ensemble with kNN as the heart of diverse sample sets. Boosting, another famous ensemble algorithm, was successfully integrated with kNN to construct an effective ensemble (García-Pedrajas and Ortiz-Boyer 2009), and was successfully applied to forecast business failure (Cortes et al. 2007, 2008; West et al. 2005). On the other hand, the combination of bagging with the CBR remains undemonstrated in its effectiveness and feasibility in achieving better performance in EOPFOF. Whether or not these two approaches can be combined to successfully improve CBR's predictive ability in EOPFOF is of interest. A complex combining scheme of ensemble makes the CBR lose the characteristic of ease of interpretation, explanation, and implementation for business

problems. The relatively low accuracy of the CBR makes its characteristic of making predictions as well suggestions unappealing. Thus, how to improve the CBR's predictive performance without losing its unique characteristics is very important and valuable. The construction of the CBR ensemble from the viewpoint of the case base issue has seldom been researched in both areas of EOPFOF and CBR, which needs to be investigated.

3 The CBR Ensemble for EOPFOF from Multiple Case Bases

3.1 Proposal of the CBR ensemble from case base issue

The case base and reasoning are two fundamental issues in CBR. The reasoning processes are founded on the case base. The so-called case base refers to historical cases related to a specific problem, which is EOPFOF in this research. Various CBR predictors are generated from one case base by using different reasoning techniques, e.g., different retrieval techniques, or by using different forms of one technique, e.g., different kNN algorithms or different case representations, both of which has been studied before. The use of different case bases has the potential ability to produce diverse CBR predictors, which has seldom been explored in the area of EOPFOF and CBR. Since the case base is the foundation of the reasoning process, research into how to produce diverse case bases could not only help the CBR produce a better predictive performance, but also help the current techniques do a better work in ensemble.

The so-called bagging provides a potential way of constructing diverse multiple case bases for CBR. Bagging refers to re-sampling a training dataset to create diverse predictors, which are further combined. The resampling process is implemented both with or without replacement. Bagging is implemented by resampling with replacement. The volume of samples in bagging is set as the same as that of original training dataset. Once the member predictors have been trained, their predictions should be aggregated by voting for symbolic ones. The diversity that is necessary to carry out ensemble work is generated by using different training datasets. In order to carry

out bagging work in ensemble, base classifiers should be unstable. However, CBR with kNN as the heart is stable. Thus, it is valuable to investigate how to help the CBR become unstable in ensemble computing.

According to the architecture of CBR, case base is a critical component. It is possible to make the reasoning technique produce different results by creating different case bases. The CBR is an imitation of human beings' actions in real life. Human beings make their decisions according to their experience. If the experience of two people are different, it is possible and reasonable that decisions of the two people will be different. This phenomenon is the reason why group decision should be widely used in society and why ensemble computing is useful. Thus, in order to make the CBR with kNN produce diverse member predictors, a potential alternative is to create different experienced bases for each member CBR. The control of sample volume in the resampling process of bagging is useful in creating diverse case bases.

3.2 Multiple case-basing for EOPFOF with the CBR ensemble

Only three *R*s of the R^4 model are useful in EOPFOF, i.e., retrieval, reuse and retaining, if parameters are predetermined. When the idea of ensemble is introduced into a CBR-based EOPFOF, the R^3 model should be further revised and developed. Notice that the *R* of revision is added into lifecycle model when the CBR is used to solve other tasks other than EOPFOF. In this research, we introduce the basic idea of bagging with proportion case-basing to create diverse case bases. Based on these multiple case bases, the algorithms of retrieval and reuse are implemented to generate diverse member CBR predictors. Then, the outputs of these member predictors are integrated to produce the final prediction. Thus, the R^3 model is revised as the M^3-E-R model, i.e., multi-case-basing, multi-case-retrievals, multi-case-reuses, ensemble and case retain, when bagging is introduced into CBR's lifecycle from the viewpoint of case base. The M^3-E-R lifecycle model of the CBR ensemble is illustrated in Fig. 1.

Figure 1 here.

The assumption in the CBR-based EOPFOF is that similar sample companies are used to predict target companies' business states, i.e., failure or non-failure. Traditionally, a single CBR predictor is used to make the prediction. The abundant use of group decision in society provides a potential way of improving this assumption. The probability value of a majority voters being correct is larger than that of each voter when the probability value of each voter being correct is larger than 0.5 (Dietterich, 2000). Assuming we have 3 member predictors: $\{p_1,p_2,p_3\}$ for a new case x. If the error rates produced by the three member predictors are uncorrelated, then p_2(x) and p_3(x) may be right when p_1(x) is wrong. A majority vote of the three member predictors will correctly predict x. Thus, the assumption of EOPFOF with the CBR ensemble is that the majority of predictions from each single CBR predictor is used to label the business states of target companies. Under this assumption, the M^3-E-R model is illustrated as follows:

- For the task of EOPFOF, initial samples and features should firstly be collected and selected. The case base is made up of experienced cases represented by significant features. The case representation issue in the traditional lifecycle of the CBR does not change in this scheme. This process is still a critical issue in the M^3-E-R model. When it is necessary to add the new target company into the case base, the process of case retaining takes place. This process is also the same as the traditional one.
- The only factor that makes the traditional R^3 model for EOPFOF different is the introduction of bagging. All changes in the lifecycle model relate to the introduction of bagging with original case base. The aim of multiple case-basing by bagging is to generate diverse case sub-bases, each of which holds some different proportion of unique experienced cases. The volume of each case base for the member CBR predictors is the same. These multiple case bases are the foundation of diverse and accurate member CBR predictors. Both diversity and accuracy of CBR predictors are a product of this scheme.

- On the foundation of various multiple case bases, the algorithms of case retrieval and case reuse are implemented, which are the kNN algorithm and majority voting respectively. Thus, various member CBR predictors implement independent retrievals of similar cases and make independent predictions by reusing solutions of similar cases. This is known as case multi-retrievals and case multi-reuses in the M^3-E-R model. One can use several other algorithms of case retrieval and case reuse in further research.
- The predictions of the member CBR predictors generated on top of diverse multiple case bases are combined to produce the final prediction. This process is called the ensemble of member CBR predictors in the M^3-E-R model. The combining scheme is majority voting according to bagging.

3.3 Control of volume in each case base

When introducing bagging to produce multiple case bases, the volume of each case base should be determined. Assuming that the number of samples in the original training dataset is expressed as M. Let M_i express the volume value of the ith resampled case base. The bagging proportion, i.e., BP, is defined as follows.

$$BP_i = \frac{M_i}{M} \in (0,1], \tag{1}$$

Bagging produces resampling datasets with the volume value being the same as the original training dataset, i.e., $M_i=M$. However, we use the parameter of BP to control the volume of each case base. Thus, we have called the revised bagging algorithm the proportion bagging for the case base issue of CBR ensemble. All multiple case bases have the same volume, i.e., $M_1=M_2=\ldots=M_i=\ldots=M_L$. The algorithm is illustrated as follows.

PROPORTION BAGGING FOR THE CASE BASE ISSUE

Training phase

1). Initialize the parameters

- D=[ ], expressing the CBR ensemble
- L, expressing the number of member CBR predictors to train.

2). For i = 1:1: L

- Take a bootstrap sample S_i, i.e., the ith case base of the ith member CBR predictor, from CB, i.e., the initial case base, with replacement; each of the multiple case bases has *ceil* ($BP_i \times M$) samples. *ceil* ( ) is function that rounds the elements in ( ) to the nearest integers towards infinity.
- Create a member CBR predictor D_i on top of S_i
- Add the member CBR predictor to the current ensemble, $D=D \cup D_i$.

3). Return D.

Prediction phase

4). Run D_1, …, D_L on the target sample company, TC

5). The class with the maximum number of votes is used to label TC.

Consider the following illustration with three *Ss*, i.e., case sub-bases from the initial case base *CB*. Each of the three case sub-bases is generated by randomly retrieving ($BP_i \times M$) samples from the *CB*. Assume that the Case Sub-base A consists of the first 10 cases. The Case Sub-base B consists of the last 10 cases, and the Case Sub-base C consists of case 3-12. Then, three CBR modulars are generated on the three Case Sub-bases. In each CBR modular, the labels of the most similar cases are used to generate a prediction of the unknown case. For example, if there are 6 cases with the label 1 and 4 cases with the label -1 in the CBR modular A, the output of the modular is 1.

According to the M^3-E-R model and algorithm of proportion bagging, the structure of the CBR ensemble based on multiple case bases generated by proportion bagging is illustrated as Fig. 2.

Figure 2 here.

3.4 Multiple case retrievals, multiple case reuses and ensemble

All the other mechanisms of the classic CBR remain unchanged except that bagging is revised to be proportion bagging and introduced to produce diverse multiple case bases. For each member CBR predictor, the classic R^4 model or the R^3 model for EOPFOF is implementable. The key issue in multiple case retrievals is to calculate similarity between a pair of cases, e.g., c_a and c_b in each member CBR predictor independently. The value of this similarity is revised from Euclidean distance, which is illustrated as follows.

$$Sim(c_a, c_b) = f(EuclideanDis(c_a, c_b)), \tag{2}$$

Multiple case reuses are used to make predictions within each member CBR predictor by simply voting for k nearest neighbors independently. This scheme is expressed as follows.

$$prediction(D_i) = sign(\sum_k label(nearestneighbor_{k,i})), \tag{3}$$

Where *prediction* ( ) means the precision of the *i*th member CBR predictor, *sign* ( ) is a function which only returns a -1 or +1, *label*( ) is a function that returns labels of the nearest neighbors. This voting scheme is also used in aggregating the results of all member CBR predictors in ensemble process, that is,

$$prediction(D) = sign(\sum_i label(D_i)), \tag{4}$$

Consider an illustration with three CBRs, namely: A, B, and C. The most similar case sets from CBRs A,B, and C are, respectively, represented by Case set A, Case set B, and Case set C. All three case sets are used to label the unknown Case D independently. Then the labels are integrated to generate a final result. If the outputs of CBR A and CBR B are 1, and the output of Case C is -1, then the label of Case D generated by the ensemble system will be 1. The CBR ensemble from multiple case-basing assumes that the random proportion case-basing process is able to generate diverse case bases from all available samples. These diverse case bases are supposed to help CBR generate specific diverse predictive modules. Finally, the integration of outputs of diverse predic-

tive modules will make more accurate prediction. Thus, the parameter of *BP* is important in assuring diverse case bases.

4 Empirical Design

The objective of this empirical research is twofold. One is to investigate empirically an appropriate parameter range of multiple case-basing with proportion bagging, i.e., *BP*, in Chinese EOPFOF prior to 3 years ago. The other one is to investigate whether the CBR ensemble from multiple case bases can be successfully used in EOPFOF of China.

4.1 Design of data analysis

We used total predictive accuracy as the assessment. Multiple splits of a dataset into a training and validating sets or a training and a testing sets was used to produce a series of accuracies, with which statistical techniques were further used to assess the predictive performance. Since feature selection methods were used to help represent the data, and parameter optimization was used to help CBR produce better performance, it is more reasonable to implement the feature selection and parameter optimization approaches on data seen and then assess the model performance on unseen data. When optimizing a method, more samples are supposed to be used. Thus, the whole available dataset was split into two parts (70% and 30%). The first part was used as a seen dataset. Feature selection and parameter optimization were implemented on the seen data. All of the seen data was used in feature selection. 30% of the seen data was used as validating set in parameter optimization, and the split was repeated many times. The other part was used as unseen data to assess model performance. In order to make the assessment significant in statistic, 30%, 50%, 70% of the unseen data was used respectively to assess the model's performance. This treatment was repeated 200 times. Thus, 200 observed results on various models were generated. Statistical indices of mean accuracy, minimum accuracy, maximum accuracy, median accuracy, standard deviation (SD), and range were generated from all observations. These statistical indices were used to

discuss predictive performances of methods. The integration of discussions on the three different data splits was used to produce the final assessment.

4.2 Data collection and representation

Business failure in China is defined as two consecutive years' negative net profit. Non-failed samples are defined as those companies which have never been specially treated. 320 non-failed samples were randomly paired with 320 failed-samples in the period of 2001-2010. The dataset was for prediction prior to 3 years ago. Published statements, including: income statement, balance sheet, and stock information from the stock exchange were used to calculate financial ratios, which were used to represent cases. 24 features, which are significant in identifying failed samples from non-failed ones from the perspective of ANOVA (Analysis of Variance), were used as initial features. Stepwise MDA was further used to filter significant ratios in EOPFOF. The procedure of the use of MDA to select feature is as follows. Firstly, all variables were assessed and ranked on their contributions in helping the MDA make prediction. The most important feature was selected. Then, the most important one in the remaining features was added one-by-one each time to generate various feature subsets. Performances of MDA on these feature subsets were assessed. Finally, the feature subset that helped the MDA produce the best performance was used. Features used to represent the cases are illustrated in Table 1. All data values were scaled into the range of [0,1] by min-max normalization.

Table 1 here.

4.3 Experimental settings

The number of nearest neighbors, i.e., k, was set as 5. The number of total member CBR predictors, i.e., L, was set as 10 according to the setting in classic bagging algorithm (Quinlan, 1996). In order to find an appropriate range of BP, this parameter was searched in the range of [0.1:1] with the step of 0.1 by using grid-search technique on the seen dataset.

5 Results and discussion

5.1 The range of *BP* and corresponding performances

In order to make the CBR ensemble with multiple case bases work well, feasible range of the parameter of proportion bagging for case base issue, i.e., *BP*, should be investigated empirically. The classic bagging algorithm is a specific form of proportion bagging with the setting of *BP* =1. The results of the search of appropriate values of *BP* are illustrated as Fig. 3.

Figure 3 here.

From Fig. 3 we see that the CBR ensemble from multiple case bases with proportion bagging produced the best performance in case representation by MDA features when the value of *BP* is 0.5. This performance is followed by those with *BP* values as 0.9, 0.4, 0.6 and 0.3 in descending sequence. This finding means that the appropriate range of *BP* for the CBR ensemble with MDA features is [0.3:0.6, 0.9]. This finding means that volumes of all multiple case bases can be controlled as 50% of the volume of original case base. This setting of *BP*, i.e., 0.5, was used in the following experiment.

Figs. 3 also indicates that predictive performance of the CBR ensemble with the *BP* value of 1, i.e., CBR ensemble from multiple case bases with bagging, produced the worst performance, i.e., 75.39%. This finding indicates that the classic bagging is the least effective and feasible process in the creation of multiple case bases for CBR ensemble with kNN as the heart. Meanwhile, this result again demonstrates the necessity and feasibility of using a volume parameter in bagging, i.e., *BP*, to control the size of multiple case bases for member CBR predictors. This finding also indicates that the use of proportion bagging to produce multiple case bases in the construction of the CBR ensemble is reasonable and necessary. The result supports the treatment of introducing a volume parameter to control the size of each of the multiple case bases. The CBR ensemble is supposed to produce a more accurate performance by using this volume parameter.

5.2 Predictive performances of the CBR ensemble and comparative models

With the *BP* value set as 0.5, the performance of the CBR ensemble based on multiple case bases under proportion bagging is compared with those of the MDA, Logit, the classic CBR and its best member predictor. The results of predictive performances of comparative methods are presented from the six statistics, i.e., mean, median, s.d., minimum, maximum, and range, as shown in Tables 2-4. Paired-samples *t* test was used to test the four hypotheses. Table 5 shows the results of significance test according to the hypotheses.

Tables 2-5 here.

From Table 2 we find that the CBR ensemble base on multiple case bases under proportion bagging with *BP*=0.5 and the proportion of training and testing sets as 70%:30% produced the largest value of mean accuracies, i.e., 77.78%. This value is better than those of the MDA, Logit, CBR and its best member predictor, respectively, by 1.21, 2.18, 0.76, and 4.03 in absolute value. This finding means that the CBR ensemble under proportion bagging can reduce error rates of comparative methods, respectively, by 5.16%, 8.93%, 3.31%, and 15.35%. Thus, in terms of mean accuracy, the CBR ensemble is better than all the comparative methods in Chinese EOPFOF prior to 3 years ago. From Table 3 we find that the CBR ensemble with *BP* as 0.5 and the proportion of training and testing sets as 50%:50% produced the largest value of mean accuracy, i.e., 77.05%. This value is better than those of MDA, Logit, CBR, and its best member predictor, respectively, by 1.10, 2.16, 0.68, and 4.11 in absolute value. The CBR ensemble based on multiple case bases under proportion bagging reduced error rates of comparative methods, respectively, by 4.57%, 8.60%, 2.88%, and 15.19%. In terms of mean accuracy, the CBR ensemble is better than comparative methods for Chinese EOPFOF prior to 3 years ago. From Table 4 we find that the CBR ensemble with *BP* as 0.5 and the proportion of training and testing sets as 30%:70% produced the second best value of mean accuracy, i.e., 74.49%. This value is superior to those of all the comparative methods, respectively, by 0.16, 1.24, and 4.43, except for the CBR. The CBR produced the best performance, i.e.,

74.58%. This result indicates that the CBR ensemble reduced the error rates of corresponding comparative methods, respectively, by 0.62%, 4.64%, and 14.80%. This finding means that all the comparative methods are at least not better than the CBR ensemble in terms of mean accuracy. We also find that MDA produced a little better performance than logit. The logit model is transferred from MDA. Assuming that the discriminating function is expressed by $f(x)$, where x is a vector. The logit function is transferred from MDA by $1/[1+\exp(-f(x))]$. The transfer did not achieve accuracy improvement for the application problem.

5.3 Significance test and analysis

In order to find whether or not there are significant differences between the CBR ensemble from multiple case bases and each comparative model in Chinese EOPFOF, paired-samples *t* test was used. If the hypothesis that no significant difference exists between each pair of compared models is rejected, we conclude that significant difference exists. Hypotheses in the null forms are presented as follows:

- H_1: Performance of the CBR ensemble is not significantly different from that of the classic CBR.
- H_2: Performance of the CBR ensemble is not significantly different from that of the MDA.
- H_3: Performance of the CBR ensemble is not significantly different from that of Logit.
- H_4: Performance of the CBR ensemble is not significantly different from that of the best member.

From results of significance test, i.e., Table 5, we find that all the four hypotheses are rejected in the mull forms on the three case representations, except that H_1 and H_3 are accepted with results produced through the use of 70% of the unseen dataset as the testing set. The significant levels are all 1%. This result means that the CBR ensemble based on multiple case bases under proportion bagging is significantly better than the MDA, Logit, and its best member CBR predictor on the whole. For the condition that H_1 and H_3 are accepted, it means that the CBR ensemble is as good as MDA and the classic CBR when 70% of the unseen dataset was used as the testing set. However, the CBR ensemble is more dominating than MDA and the classic CBR in terms of the three types

of data splits. This assertion is supported by Table 6, which is a voting conclusion from empirical results of the three types of experiment. From Table 6 we find that voting results of various predictive methods on experiments of the three types of data split indicate that the CBR ensemble based on multiple case bases under proportion bagging is significantly better than the MDA, Logit, the classic CBR, its best member CBR predictor in terms of predictive performances of Chinese EOPFOF prior to 3 years ago.

Table 6 here.

In general, we can assert that the CBR ensemble based on multiple case bases under proportion bagging with *BP* value as 0.5 can produce a significantly better performance than the MDA, Logit, CBR and its best member predictor from the views of the three types of data splits, i.e., the proportion of training set and testing set from unseen data, respectively, as 70%:30%, 50%:50%, and 30%:70%. The dominating performance of the CBR ensemble attributes to the introduction of a volume parameter to bagging and the use of this revised algorithm to generate diverse multiple case bases for all member CBR predictors. Thus, diverse and accurate member CBR predictors are produced, on top of which, an ensemble of CBR is generated. The CBR ensemble based on multiple case bases from this scheme not only produces more dominating performance than CBR ensemble under bagging, but also is superior to its best member CBR predictor, the classic CBR and the two famous statistical methods of MDA and Logit. These results demonstrate the effectiveness and feasibility of our proposal to construct the CBR ensemble for EOPFOF. This research presents a new perspective into improving the CBR's predictive performance by focusing on the case base issue. We obtained the objective of improving the CBR's predictive performance by using techniques in the case base.

5.4 Implication and contribution

The implications of this research are three-fold. 1) The results of this research indicate that the CBR ensemble from multiple case bases has the potential of improving the predictive performance of business failure. Former

research into the CBR or CBR-based EOPFOF chiefly focuses on optimizing the single CBR to improve its performance. They seldom attempt to integrate ensemble computing with the CBR to generate a new type of methodology from the perspective of its lifecycle model. This research made an early investigation into the integration of ensemble computing with the R^4 lifecycle model of the CBR and applied the new type of CBR methodology in forecasting business failure in China. It implicates that group decision of the CBR can take advantage of different techniques, and provides a reference for further research into the CBR ensemble. A new topic in areas of EOPFOF and the CBR emerges. 2) Previous research indicates that combination of bagging with kNN is not very useful in improving predictive performance. This research attempts to introduce a volume parameter into bagging to control the size of each case base of member predictors. This mechanism makes the CBR with kNN as the heart unstable by building up diverse case bases. It is a feasible way of helping build up more accurate CBR predictors with kNN as the heart by proportion bagging. 3) The practice implication is that multiple CBR methodology is more effective than single CBR in helping people predict business failure. This finding recommends practitioners to take into consideration of various expertises in order to make more precise decisions. Thus, the group decision of multiple techniques is a valuable topic in terms of both theory and practice, and it should be researched together with ensemble computing.

Contributions of this research are three-fold. 1) The theory contribution for the CBR is early research to build up the CBR ensemble from the case base. The ensamble of multiple CBR predictors from multiple case bases is capable of taking advantage of member predictors, which is supposed to help the CBR become more applicable. 2) The theory contribution for bagging is the introduction of the volume control parameter which makes bagging more applicable. Bagging can be effectively integrated with the CBR with kNN as the heart to solve classification or prediction problems. 3) The application contribution is early research into solving the problem of

EOPFOF together with the integration of bagging, proportion case-basing and CBR. It provides an effective alternative for EOPFOF.

6 Conclusion and Limitations

The conclusion of this research is that the CBR ensemble based on multiple case bases under proportion bagging is able to provide dominating predictive performance in Chinese EOPFOF prior to 3 years ago. The attempt at improving CBR's predictive performance by ensemble from case base issues is achieved. This scheme of introducing a parameter to control the volume of each case base of member CBR predictors is effective and feasible in ensuring that member predictors can produce diverse and accurate performances. The result of this CBR ensemble is significantly better than that from the CBR ensemble with bagging, MDA, Logit, single CBR, and the best member CBR predictor. We used three types of data splits to support this experiment, namely: the use of 30%, 50%, and 70% of unseen data as testing set. Thus, the scheme was demonstrated to be effective and feasible in combing bagging with the CBR. This scheme made CBR retain the characteristic of interpretation, and obtain an improvement in predictive performance. Consequently, the CBR methodology will play a more important role in forecasting business failure.

This research has the following limitations, on which further investigations should focus. This research concentrates on the prediction of business failure in China, since China is one of the most quickly developing countries in the world. Many companies in western developed countries have set up manufacturing factories in China. Thus, the study of failure possibility evaluation of firms in China is of interest to both Chinese people and western people. The conclusion of this research is that integration of CBR with ensemble learning has the potential to improve the predictive performance of business failure. This method has the potential to help predict western business failure. Further research should be carried out to demonstrate the effectiveness and feasibility of this new CBR ensemble from multiple case bases under proportion bagging with data collected from western coun-

tries for EOPFOF. What needs to be done includes: the collection of a dataset, the selection of significant financial ratios, and the optimization of the method for western EOPFOF. In order to implement cross-cultural research, it is valuable to collect some western datasets for this task in the following work. The investigation on the use of this CBR ensemble in tackling similar problems is also valuable, e.g., credit scoring, pattern recognition. This research has demonstrated the effectiveness and feasibility of focusing on the case base when constructing the CBR ensemble. Further research can also be carried out to explore the CBR ensemble from other aspects.

Acknowledgements

This research is partially supported by the National Natural Science Foundation of China (No. 70801055) and the Zhejiang Provincial Natural Science Foundation of China (No. Y7100008). The authors gratefully thank anonymous referees for their useful comments and editors for their work.

References

[1] A. Aamodt, E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approach. AI Communications 7 (1) (1994) 39-59.

[2] H. Ahn, K. Kim, Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach. Applied Soft Computing 9(2) (2009) 599-607

[3] E. Alpaydin. Voting over multiple condensed nearest neighbors. Artificial Intelligence Review 11 (1997) 115–132.

[4] H. Altınçay. Ensembling evidential k-nearest neighbor classifiers through multi-modal perturbation. Applied Soft Computing 7(3) (2007) 1072-1083

[5] Y. Bao, N. Ishii. Combining multiple k-nearest neighbor classifiers for text classification by reducts. In S. Lange, K. Satoh and C. Smith (Eds.), Proceedings of the Fifth International Conference on Discovery Science, vol. 2534 of Lecture Notes in Computer Science, 2002, pp. 340–347

[6] Y. Bao, N. Ishii, X. Du. combining multiple k-nearest neighbor classifiers using different distance functions. In Z. Yang, R. Everson and H. Yin (Eds.), Proceedings of the Fifth International Conference on Intelligent Data Engineering and Automated Learning, vol. 3177 of Lecture Notes in Computer Science, 2004, pp. 634–641

[7] S. Bay. Combining nearest neighbor classifiers through multiple feature subsets. In J. Shavlik (Ed.), Proceedings of the Fifteenth International Conference on Machine Learning, Madison, WI, 1998, pp. 37– 45

[8] G.R. Beddoe, S. Petrovic, Selecting and weighting features using a genetic algorithm in a case-based reasoning approach to personnel rostering. European Journal of Operational Research 175(2)(2006) 649-671

[9] L. Breiman. Bagging predictors. Machine Learning 24 (1996) 123-140.

[10] L. Breiman. Pasting small votes for classification in large databases and on-line. Machine Learning 36 (1999) 85–103.

[11] J.L. Castro, M. Navarro, J.M. Sánchez, J.M. Zurita. Loss and gain functions for CBR retrieval. Information Sciences 179(11) (2009) 1738-1750.

[12] P.-C. Chang, L.-Y. Chen, C.-Y. Fan, A case-based evolutionary model for defect classification of printed circuit board images. Journal of Intelligent Manufacturing 19 (2008) 203-214

[13] P.-C. Chang, C.-Y. Lai, K.R. Lai, A hybrid system by evolving case-based reasoning with genetic algorithm in wholesaler's returning book forecasting. Decision Support Systems 42(3) (2006) 1715-1729

[14] P.-C. Chang, C.-H. Liu, K.R. Lai, A fuzzy case-based reasoning model for sales forecasting in print circuit board industries. Expert Systems with Applications 34 (2008) 2049-2058.

[15] E. Cortes, M. Martınez, N. Rubio. A boosting approach for corporate failure prediction. Applied Intelligence 27(2007)29–37

[16] E. Cortes, N. Rubio, M. Martınez, D. Elizondo. Bankruptcy forecasting: An empirical comparison of Ada-

Boost and neural network. Decision Support System 45 (1) (2008) 110-122

[17] P. Cunningham, G. Zenobi. Case representation issues for case-based reasoning from ensemble research. In: ICCBR, 2001, pp. 146-157

[18] T.G. Dietterich. Ensemble methods in machine learning. Lecture Notes in Computer Science 1857 (2000) 1-15.

[19] C.-Y. Fan, P.-C. Chang, J.-J. Lin, J.-C. Hsieh. A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Applied Soft Computing 1(1) (2011) 632-644

[20] N. García-Pedrajas, D. Ortiz-Boyer. Boosting k-nearest neighbor classifier by means of input space projection. Expert Systems with Applications 36(7) (2009) 10570-10582

[21] G.H. Lee, Rule-based and case-based reasoning approach for internal audit of bank. Knowledge-Based Systems 21(2) (2008) 140-147

[22] H. Li, J. Sun. Ranking-order case-based reasoning for financial distress prediction. Knowledge-Based Systems 21 (8) (2008) 868-878.

[23] H. Li, J. Sun. Predicting business failure using multiple case-based reasoning combined with support vector machine. Expert Systems with Applications 36 (6) (2009) 10085-10096

[24] H. Li, J. Sun. Forecasting business failure in China using case-based reasoning with hybrid case respresentation. Journal of Forecasting 29 (6) (2010) 558-573.

[25] R. Lin, Y. Wang, C. Wu, et al., Developing a model for failure possibility evaluation of firms via RST, GRA and CBR. Expert Systems with Applications 36 (2) (2009) 1593-1600.

[26] C.-H. Liu, L.-S. Chen, C.-C. Hsu, An association-based case reduction technique for case-based reasoning. Information Sciences 178 (17) (2008) 3347-3355.

[27] O. Okun, H. Priisalu. Multiple views in ensembles of nearest neighbor classifiers. In Proceedings of the

Workshop on Learning with Multiple Views, 22nd ICML, Bonn, Germany, 2005.

[28] S.K. Pal, S.C.K. Shiu, Foundations of Soft Case-Based Reasoning, Wiley & Sons, New Jersey (2004).

[29] C. Park. I. Han, A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction. Expert Systems with Applications 23 (3) (2002) 255-264.

[30] JR Quinlan. Bagging, Boosting, and C4.5. In: Proceedings of the Thirteenth National Conference on Artificial Intelligence, 1996, pp. 725-730.

[31] M. Salamo. M. Lopez-Sanchez, Rough set based approaches to feature selection for case-based reasoning classifiers. Pattern Recognition Letters 32(2)(2011) 280-292.

[32] D. Skalak. Prototype selection for composite nearest neighbor classifiers. PhD thesis, Department of Computer Science, University of Massachusetts

[33] J. Sun, X. Hui, Financial distress prediction based on similarity weighted voting CBR. In: Li. X. et al. (Eds.). Advanced Data Mining and Applications. Springer. Berlin. 2006, pp. 947-958.

[34] D. West. S. Dellana. J. Qian, Neural network ensemble strategies for financial decision applications. Computers & Operations Research 32(10)(2005) 2543-2559

[35] D. Yankow, D. DeCoste, E. Keogh. Ensembles of nearest neighbor forecasts. 17th European Conference on Machine Learning, Proceedings. Lecture Notes in Computer Science, 2006, pp. 545-556

[36] A.Y.N. Yip, Predicting business failure with a case-based reasoning approach, In: Negoita, M., Howlett, R., Jain, L., et al. (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Springer-Verlag, Berlin, 2004, pp. 665-671.

[37] Z.-H. Zhou, Y. Yu. Adapting bagging to nearest neighbor classifiers. Journal of Computer Science and Technology 20 (2005) 48-54.

[38] Z.Y. Zhuang, L. Churilov, F. Burstein, K. Sikaris, Combining data mining and case-based reasoning for

intelligent decision support for pathology ordering by general practitioners. European Journal of Operational Research 195(3)(2009) 662-675

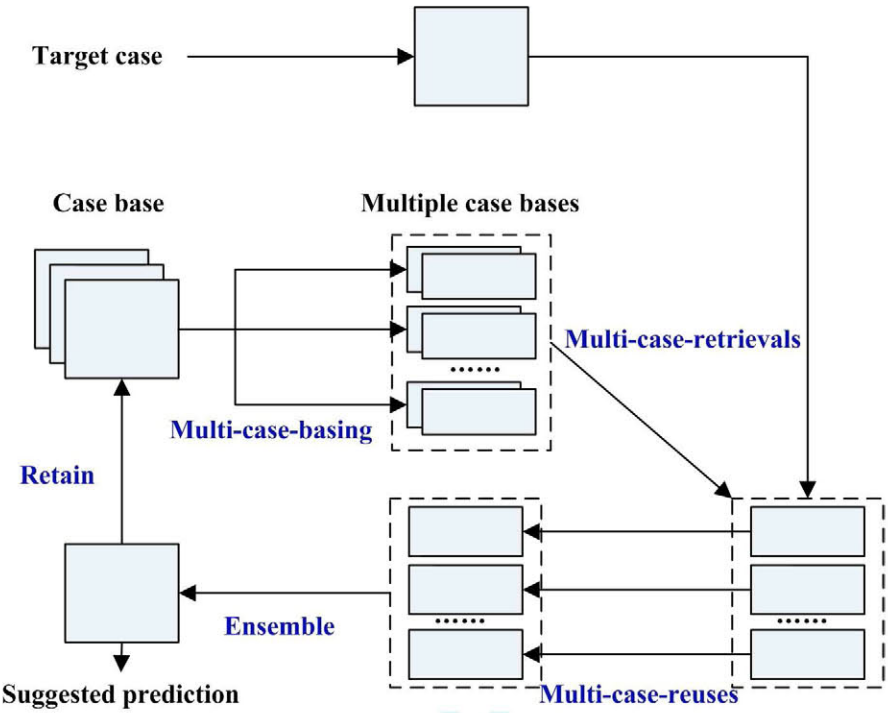

Fig. 1. The lifecycle model of the CBR after integrating proportion bagging with the CBR

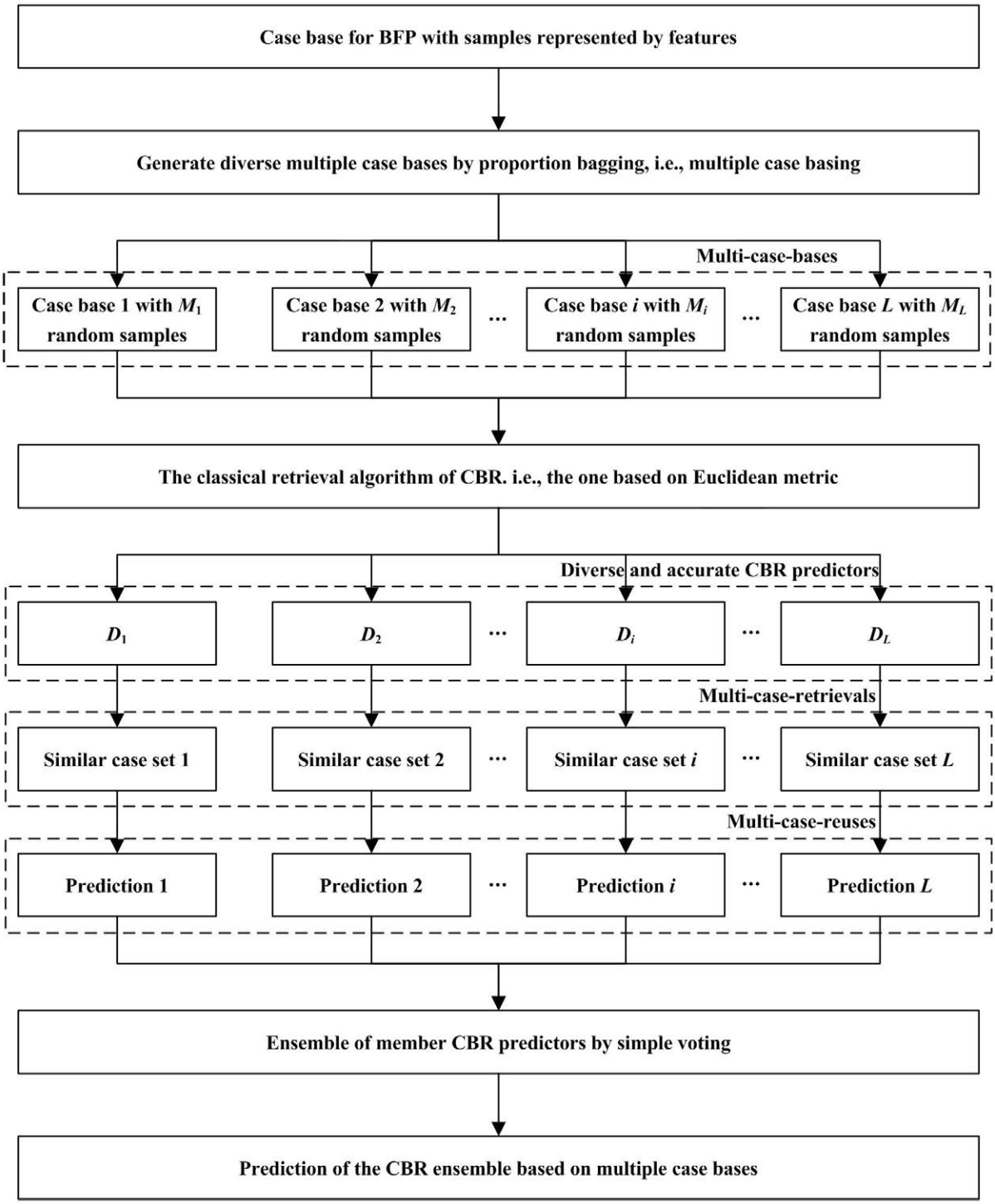

Fig. 2. Structure of the CBR ensemble based on proportion bagging for EOPFOF

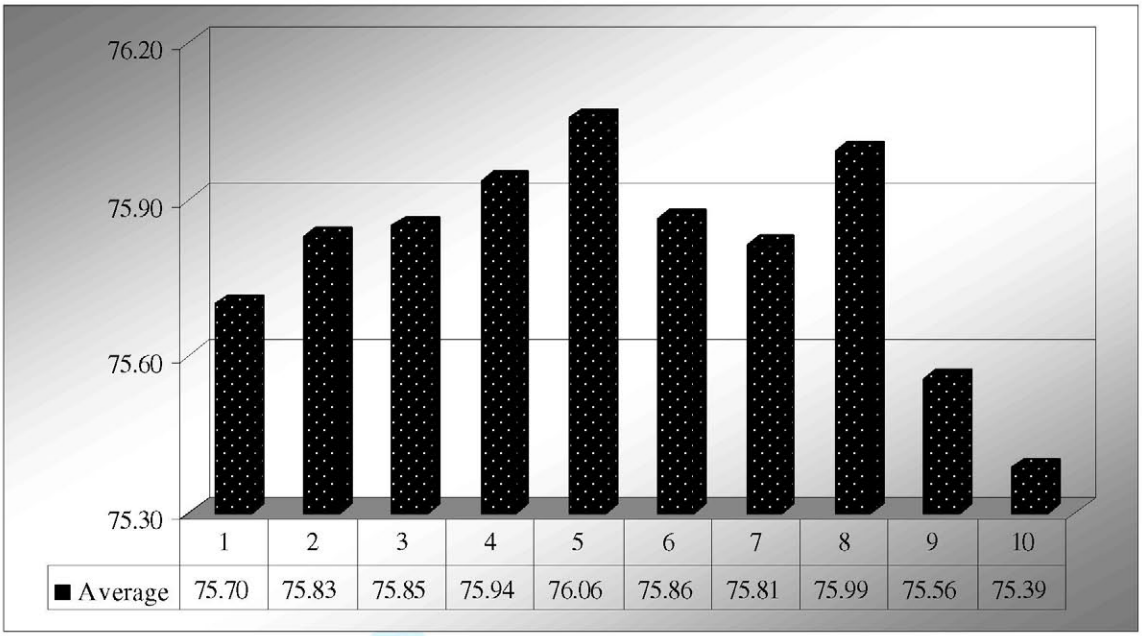

Fig. 3. Performance of the CBR ensemble with various values of *BP* on case representation with MDA features

Table 1. Three case representations for Chinese EOPFOF

No.	Feature name	Case representation using MDA features
1	Working capital ratio	
2	Capital intensity	
3	Total asset turnover	X
4	Equity to liabilities	
5	Return of assets	
6	Net income on total assets	
7	Operating income to EBT	X
8	Earning per share	
9	Net asset per share	X
10	Price per book value	X
11	Operating income per share	
12	Price-to-sales ratio	
13	Retained earnings to assets	
14	Profitability ratio per share	X
15	Market value	
16	Tobin's Q	
17	Book-to-market ratio	
18	Surplus reserves per share	
19	Inappropriate profit per share	
20	Retained earnings per share	X
21	Net cash flow from operating activities to total assets	X
22	Net cash flow from operating activities per share	
23	Net cash flow from investment activities per share	
24	Net cash flow per share	

Table 2. Statistics of predictive performances with the proportion of training set and testing set as 70% to 30%

	Methods	Mean	Median	SD	Minimum	Maximum	Range
Comparative methods	MDA	76.57	76.67	4.81	**65.00**	88.33	**23.33**
	Logit	75.60	75.00	5.14	61.67	88.33	26.67
	CBR	77.02	76.67	**4.54**	63.33	88.33	25.00
The best member	The 10th member	73.75	75.00	5.44	56.67	86.67	30.00
The CBR Ensemble	CBRE	**77.78**	**78.33**	4.59	**65.00**	**90.00**	25.00

Table 3. Statistics of predictive performances with the proportion of training set and testing set as 50% to 50%

	Methods	Mean	Median	SD	Minimum	Maximum	Range
Comparative methods	MDA	75.95	76.00	3.66	66.00	84.00	18.00
	Logit	74.89	75.00	3.50	63.00	84.00	21.00
	CBR	76.37	**77.00**	**3.24**	**68.00**	**85.00**	**17.00**
The best member	The 7th member	72.94	73.00	4.86	48.00	83.00	35.00
The CBR Ensemble	CBRE	**77.05**	**77.00**	3.61	65.00	**85.00**	20.00

Table 4. Statistics of predictive performances with the proportion of training set and testing set as 30% to 70%

	Methods	Mean	Median	SD	Minimum	Maximum	Range
Comparative methods	MDA	74.33	74.29	**3.24**	**65.00**	82.14	17.14
	Logit	73.25	73.57	3.37	64.29	80.71	**16.43**
	CBR	**74.58**	**75.00**	3.40	58.57	**82.86**	24.29
The best member	The 3rd member	70.06	70.71	5.84	49.29	80.71	31.43
The CBR Ensemble	CBRE	74.49	**75.00**	4.07	63.57	**82.86**	19.29

Table 5. Results of significance test

Proportion of training set to testing set	Methods	Mean	*t* statistic and *p* value	Significant level	Hypothesis
70%:30%	CBRE	**77.78**	3.732(0.000)***	1%	Reject H_1
	MDA	76.57			
	CBRE	**77.78**	6.065(0.000)***	1%	Reject H_2
	Logit	75.60			
	CBRE	**77.78**	2.701(0.008)***	1%	Reject H_3
	CBR	77.02			
	CBRE	**77.78**	10.448(0.000)***	1%	Reject H_4
	The best member CBR	73.75			
50%:50%	CBRE	**77.05**	4.091(0.000)***	1%	Reject H_1
	MDA	75.95			
	CBRE	**77.05**	7.516(0.000)***	1%	Reject H_2
	Logit	74.89			
	CBRE	**77.05**	2.733(0.007)***	1%	Reject H_3
	CBR	76.37			
	CBRE	**77.05**	11.826(0.000)***	1%	Reject H_4
	The best member CBR	72.94			
30%:70%	CBRE	74.49	0.446(0.656)	-	*Accept* H_1
	MDA	74.33			
	CBRE	74.49	3.521(0.000)***	1%	Reject H_2
	Logit	73.25			
	CBRE	74.49	-0.338(0.736)	-	*Accept* H_3
	CBR	**74.58**			
	CBRE	74.49	9.662(0.000)***	1%	Reject H_4
	The best member CBR	70.06			

Table 6. Voting results from the three types of data split

Method	Training set to testing set	The feature-bagging-based CBR ensemble	Voting result
MDA	70%:30%	CBRE is significantly BETTER	CBRE is BETTER
	50%:50%	CBRE is significantly BETTER	
	30%:70%	CBRE is AS GOOD AS the method	
Logit	70%:30%	CBRE is significantly BETTER	CBRE is BETTER
	50%:50%	CBRE is significantly BETTER	
	30%:70%	CBRE is significantly BETTER	
CBR	70%:30%	CBRE is significantly BETTER	CBRE is BETTER
	50%:50%	CBRE is significantly BETTER	
	30%:70%	CBRE is AS GOOD AS the method	
The best member CBR	70%:30%	CBRE is significantly BETTER	CBRE is BETTER
	50%:50%	CBRE is significantly BETTER	
	30%:70%	CBRE is significantly BETTER	