Neural Comput & Applic (2017) 28 (Suppl 1):S41–S50
DOI 10.1007/s00521-016-2304-x

ORIGINAL ARTICLE

Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance

You Zhu[1] · Chi Xie[1,2] · Gang-Jin Wang[1,2] · Xin-Guo Yan[1]

Received: 26 October 2015 / Accepted: 30 March 2016 / Published online: 19 April 2016

Abstract Supply chain finance (SCF) becomes more important for small- and medium-sized enterprises (SMEs) due to global credit crunch, supply chain financing woes and tightening credit criteria for corporate lending. Currently, predicting SME credit risk is significant for guaranteeing SCF in smooth operation. In this paper, we apply six methods, i.e., one individual machine learning (IML, i.e., decision tree) method, three ensemble machine learning methods [EML, i.e., bagging, boosting, and random subspace (RS)], and two integrated ensemble machine learning methods (IEML, i.e., RS–boosting and multi-boosting), to predict SMEs credit risk in SCF and compare the effectiveness and feasibility of six methods. In the experiment, we choose the quarterly financial and non-financial data of 48 listed SMEs from Small and Medium Enterprise Board of Shenzhen Stock Exchange, six listed core enterprises (CEs) from Shanghai Stock Exchange and three listed CEs from Shenzhen Stock Exchange during the period of 2012–2013 as the empirical samples. Experimental results reveal that the IEML methods acquire better performance than IML and EML method. In particular, RS–boosting is the best method to predict SMEs credit risk among six methods.

✉ Chi Xie
xiechi@hnu.edu.cn

[1] College of Business Administration, Hunan University, Changsha 410082, China

[2] Center of Finance and Investment Management, Hunan University, Changsha 410082, China

Keywords Supply chain finance · Credit risk · Small- and medium-sized enterprises · Core enterprises · Individual machine learning · Ensemble machine learning · Integrated ensemble machine learning

1 Introduction

In recent years, small- and medium-sized enterprises (SMEs) are confronted with huge challenges such as high operational risks, capital constrained and financial information opaque. The essential problem is that SMEs are still squeezed out of conventional financing patterns even though the policies of the Chinese government are quite helpful. As a result, the solutions of SMEs' financing issues today focus on the design and optimization of the finance pattern. As a means of substituting for lower credit availability, supply chain financing (SCF) is increasingly generating much enthusiasm among SMEs, relevant CEs and relevant financial institutions. Because SCF is strategic for overcoming SMEs' financing difficulty, ensuring CEs' continuity of production and financial institutions making correct credit loan decision. SCF can be defined a kind of financing pattern that managing, planning and controlling all the transaction activities and processes related to the flow of cash among supply chain stakeholders in order to improve turnover efficiency of working capital [1]. Roughly speaking, in SCF, the financial institutions dramatically increase their incomes; the SMEs obtain loans from financial institutions with looser constraints; and core enterprises (CEs) also alleviate the pressure of funding [2–4]. Nevertheless, SCF is unable to completely eliminate credit risk which is still one of the major threats to credit-granting institutions [5–7].

It is well known that a fraction of improvement in prediction accuracy also significantly eliminates credit risk for financial institutions [8, 9]. Therefore, predicting SMEs credit risk is a hot topic due to its importance for financial institutions making correct credit loan decision in SCF. Prediction models of corporate credit risk are developed to classify loan customers as either a credit non-risk group or a credit risk group. The functions of prediction models include reducing the cost of credit analysis, enabling fast credit decisions, close monitoring of existing accounts, and prioritizing collections [10]. To predict dichotomous outcomes of good and bad credit classes in the financing market, some traditional statistical methods and machine learning methods (a field emerges from artificial intelligence) are widely applied, such as logistic regression analysis (LRA) [11] and decision tree (DT) [12].

In research of SCF, we focus on increasing the prediction accuracy of SMEs credit risk, because a fraction of improvement in SMEs credit risk accuracy may translate into financial institutions' noteworthy future savings. Researchers consider that individual machine learning (IML) methods are prominent to the traditional statistical methods in dealing with corporate credit risk prediction issues [13]. Nevertheless, some ensemble machine learning (EML) methods show higher accuracy of predictability than any IML method, especially when the different structures of machine learning methods lead to independent errors [9, 14, 15]. Moreover, integrated ensemble machine learning (IEML) method (e.g., RS–boosting [8], the multi-boosting [16]) is a kind of powerful analytical tool which is used for further improving the prediction accuracy of SMEs credit risk by scholars. The RS–boosting is integrated by EML methods of boosting and random subspace (RS), which combine with the advantages of instance partitioning method and attribute partitioning method. The multi-boosting is integrated by EML methods of boosting and wagging, which is able to harness both boosting's high bias and variance reduction with wagging's superior variance reduction. Although above intuitive explanations are reasonable, they lack reliable experimental analysis of prediction of SMEs credit risk in SCF.

Therefore, this paper includes two research purposes: first, we prove that the two IEML methods (i.e., RS–boosting and multi-boosting) are remarkably better than IML method and EML methods in predicting SMEs credit risk by experimental analysis; second, we try to select the better one from two IEML methods by experimental analysis, which is more accuracy prediction of SMEs credit risk. The contributions of this paper are summed up as follows: (1) Our findings show that SCF is unable to completely eliminate credit risk which is still one of threats to credit-granting institutions; (2) Our primary empirical results show that the performance of DT as the base classifier is better than that of neural networks (NN); (3) Compared to previous works on evaluating results of machine learning experiments that is using type I and II errors and ROC, the proposed evaluation criteria of 'precision' rate, 'recall' rate and 'F-Measure' rate are also important for measuring the prediction performance; (4) Our empirical outcomes show that the credit risks prediction accuracy of EML method is not always better than that of IML, and RS–boosting exhibits the strongest ability of SMEs credit risk prediction in SCF than that of other five methods; (5) The integrated EML can provide a new angle of view for improving the prediction performance of machine learning, and evaluating the China's SMEs credit risk in SCF. Overall, in practical terms, our proposed IEML methods can be applied in credit risks prediction in SCF for financial institutions.

The rest of the paper is organized as follows. Section 2 discusses the methodology. Section 3 presents the details of data preparation and experiment design. The empirical results and some relevant discussions are showed in Sect. 4. Finally, some conclusions are drawn in Sect. 5.

2 Methodology

Machine learning evolves the study of pattern recognition and computational learning theory in artificial intelligence, which can learn from and make predictions on data. EML is a machine learning paradigm where multiple learners are trained to solve the same problem. In contrast to ordinary IML methods which try to learn one hypothesis from training data, EML methods try to construct a set of hypotheses and combine them [17]. Meanwhile, EML methods are noted in virtue of strong generalization ability. Nonetheless, the prediction ability of IEML methods is usually much stronger than that of a single EML method, which makes IEML methods very attractive. In the next six subsections, we firstly introduce a IML method (i.e., DT) which is used as base classifier of IEML methods in this study; then show three popular and representative EML methods, i.e., bagging, boosting and RS; finally exhibit two IEML method, i.e., RS–boosting and multi-boosting.

2.1 Decision tree (DT)

DT includes nodes, branches and leaves, which is widely applied in predicting credit risk. Nodes in a decision tree correspond to features, branches of the tree correspond to their associated values, and leaves of the tree correspond to classes. There are many specific DT algorithms such as CART, CHAID, MARS, ID3 and C4.5. In this paper, we apply the C4.5 algorithm which summarizes the training data in the form of a DT and is based on improving ID3 algorithm. C4.5 algorithm chooses the attribute that gains the most information to be at the root of the tree, which is applied recursively to form sub-trees, terminating when a

```
Input: an attribute-valued dataset $D$
1. $Tree = \{\}$
2. if $D$ is 'true' or other stopping criteria met then
3. terminate
4. end if
5. for all attribute $a \in D$ do
6. compute information-theoretic criteria if we split on $a$
7. end for
8. $a_{best}$ = Best attribute according to above computed criteria
9. $Tree$ = Create a decision node that tests $a_{best}$ in the root
10. $D_v$ = Induced sub-datasets from $D$ based on $a_{best}$
11. for all $D_v$ do
12. $Tree_v = C4.5(D_v)$
13. Attach $Tree_v$ to the corresponding branch of $Tree$
14. end for
15. return $Tree$
```

Fig. 1 The C4.5 algorithm of decision tree [19]

given subset contains instances of only one class [18, 19]. It is noteworthy that we also use DT as base classifier of IEML methods in our experiments according to Maclin and Opitz [20], Fu et al. [21] and Wang and Ma [8]. Additionally, Maclin and Opitz [20] find that the performance of DT as the base classifier is better than that of NN. In this paper, we evaluate two IEML methods on data sets using both NN (i.e., Multilayer Perceptron) and DT (i.e., C4.5) as the base classifier for choosing the appropriate one. The pseudocode of C4.5 algorithm of DT is shown in Fig. 1 [19].

2.2 Bagging

The bagging is a method for generating multiple versions of predictor and using these to obtain an aggregated predictor [22]. Beriman [22] considers that bagging demonstrates remarkable consistency in its ability to reduce error and give substantial gains in accuracy. Bagging can be employed with any base classification technique. Furthermore, bagging operate by selectively re-sampling from the training data to generate derived training sets to which the base learner is applied [16]. The pseudocode of bagging is shown in Fig. 2 [23].

2.3 Boosting

The weak learning algorithm means a little bit better than random guessing, but Boosting can be used to significantly reduce the error of any learning algorithm and is a general method to improve the performance of some weak learning algorithms [24]. There is an improved version of boosting, which is called AdaBoost. Freund and Schapire [24] argue that AdaBoost has some properties which make it more practical and easier to implement than the legacy version (i.e., boosting). They also propose another improved version based on AdaBoost, i.e., AdaBoost.M1 [24]. More precisely, in this paper we apply the AdaBoost.M1 with three reasons: (1) It uses one step weight update process that is less subject to numeric underflow than the original two-step process [16]; (2) It prevents numeric underflow [16]; and (3) It continues producing more committee members beyond the point when $\varepsilon_t > 1/2$, which is claimed to improve predictive accuracy [23]. The pseudocode of AdaBoost.M1 is shown in Fig. 3 [24].

```
Input: training set $S$ , Inducer $I$ , integer $T$ (number of bootstrap samples)
1. for $i = 1$ to $T${
2. $S'$ = bootstrap sample from $S$ (i.i.d. sample with replacement)
3.$C_i = I(S')$
4.}
5.$C^*(x) = \arg\max_{y \in Y} \sum_{i:C_i(x)=y} 1$ (the most often predicted label $y$ )
Output: classifier $C^*$
```

Fig. 2 The bagging algorithm [23]

```
Input: sequence of $m$ examples $\langle(x_1, y_1), ..., (x_m, y_m)\rangle$ with labels $y_i \in Y = \{1, ...k\}$
weak learning algorithm WeakLearn integer $T$ specifying number of iterations
Initialize $D_1(i) = 1/m$ for all $i$
Do for $t = 1, 2, ..., T$
1. Call WeakLearn, providing it with the distribution $D_t$
2. Get back a hypothesis $h_t : X \rightarrow Y$
3. Calculate the error of $h_t : \varepsilon_t = \sum_{i:h_t(x_i) \neq y_i} D_t(i)$. If $\varepsilon_t > 1/2$ , then set $T = t - 1$
and abort loop
4. Set $\beta_t = \varepsilon_t/(1 - \varepsilon_t)$
5. Update distribution $D_t : D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} \beta_t \ if \ h_t(x_i) = y_i \\ 1 \quad otherwise \end{cases}$ Where $Z_t$ is
normalization constant (chosen so that $D_{t+1}$ will be a distribution)
Output the final hypothesis:$h_{fin}(x) = \arg\max_{y \in Y} \sum_{t:h_t(x)=y} \log \frac{1}{\beta_t}$
```

Fig. 3 The AdaBoost.M1 algorithm [24]

2.4 Random subspace (RS)

In order to avoid over-fitting a set of training data while achieving maximum accuracy when use DT method. Ho [25] proposes a method for systematic construction of decision forest, which relies on a pseudorandom procedure to select components of a feature vector, and DT is generated by using only the selected feature components. This method is called RS method which is parallel learning algorithm; in other words, the generation of each DT is independent. Therefore, Ho [25] considers that RS is suitable for parallel implementation for fast learning that is desirable in some practical applications; moreover, there is no danger of being trapped in local optima. The pseudocode of RS is shown in Fig. 4 [8].

2.5 RS–boosting

RS–boosting is a kind of IEML method, which is based on two popular ensemble strategies, i.e., boosting and RS. The base classifiers of RS–boosting are trained by RS method, after that, these base classifiers are used to reweight the instances. Meanwhile, RS–boosting introduces RS strategy into each boosting iteration and combines two EML methods' advantages. In this paper, we use C4.5 as the base

Fig. 4 The Random subspace algorithm [8]

```
Input: Data set $D = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ , Base classifier algorithm $L$ ,
Number of random subspace rate $k$ , Number of learning rounds $T$
1. For $t = 1, 2, ..., T$
2. Random generate a subspace sample from $D$, $D_t = RS(D, k)$
3. Train a base classifier $h_t$ from the subspace sample, $h_t = L(D_t)$
4. end
Output: $H(X) = \arg\max_{y \in Y} \sum_{t=1}^{T} 1(y = h_t(x))$;
$\begin{cases} 1(\alpha) = 1 \text{ if } \alpha \text{ is ture} \\ 1(\alpha) = 0 \text{ otherwise} \end{cases}$
```

```
Input: Data set $D = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ , Base classifier algorithm $L$ ,
Number of random subspace rate $k$ , Number of learning rounds $T_B$, Number of learning
rounds for Random Subspace $T_R$
1.Initialize the weight distribution $D_1(i) = 1/m$
2. For $t = 1, 2, ..., T_B$
3. For $s = 1, 2, ..., T_R$
4. Random generate a subspace sample from $D_t : D_t^s = RS(D_t, k)$
Train a base classifier $h_s$ from $D$ using distribution $D_t : h_s = DT(D_t^s)$
5. end
6.$H_t^{RS}(x) = \arg\max_{y \in Y} \sum_{s=1}^{T_R} 1(y = h_s(x))$
7. Measure the error of $H_t^{RS} : \varepsilon_t = Pr_{i \sim D_i}[H_t^{RS}(x_i \neq y_i]$
Determine the weight of $H_t^{RS} : \alpha_t = \frac{1}{2} \ln \frac{1-\varepsilon_t}{\varepsilon_t}$
8. Update the distribution, where $Z_t$ is a normalization factor with enables $D_{t+1}$ to be
distribution:$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} \exp(-\alpha_t) \text{ if } H_t^{RS}(x_i) = y_i \\ \exp(\alpha_t) \text{ if } H_t^{RS}(x_i) \neq y_i \end{cases} = \frac{D_t(i) \exp(-\alpha_t y_i H_t^{RS}(x_i))}{Z_t}$
9. end
Output:$H^{RS-Boosting}(x) = sign(f(x)) = sign \sum_{t=1}^{T} \alpha_t H_t^{RS}(x)$
```

Fig. 5 The RS–boosting algorithm [8]

```
Input: $D$ , a sequence of $m$ labeled examples $\langle (x_1, y_1), ...(x_m, y_m) \rangle$ with labels
$y_i \in Y$, Base learning algorithm $L$ ,integer $T$ specifying the number of iterations, vector
of integers $I_i$ specifying the iteration at which each subcommittee $i \geq 1$ should terminate
1. $D' = D$ with instance weights assigned to be 1
2. set $k = 1$
3. For $t = 1$ to $T\{$
4. If $I_k = t$ then reset $D'$ to random weights drawn from the continuous Poisson
distribution, standardize $D'$ to sum to $n$ , increment $k$
5.$C_t = L(D')$
6.$\varepsilon_t = \frac{\sum_{x_j \in D' : C_t(x_j) \neq y_j} wight(x_j)}{m}$
7. if $\varepsilon_t > 0.5$ then reset $D'$ to random weights drawn from the continuous Poisson
distribution, standardize $D'$ to sum to $n$ , increment $k$
8. go to Step 5
9. otherwise if $\varepsilon_t = 0$ then set $\beta_t$ to $10^{-10}$ , reset $D'$ to random weights drawn from the
continuous Poisson distribution, standardize $D'$ to sum to $n$ , increment $k$
10. otherwise $\beta_t = \frac{\varepsilon_t}{(1-\varepsilon_t)}$
11. For each $x_j \in D'$
12. divide $weight(x_j)$ by $2\varepsilon_t$ if $C_t(x_j) \neq y_j$ and $2(1 - \varepsilon_t)$ otherwise
13. if $weight(x_j) < 10^{-8}$ , set $weight(x_j)$ to $10^{-8}$
14.}
Output the final classifier:$C^*(x) = \arg\max_{y \in Y} \sum_{t : C_t(x) = y} \log \frac{1}{\beta_t}$
```

Fig. 6 The multi-boosting algorithm [16]

learning algorithm of RS–boosting. Wang and Ma [8] prove that RS–boosting gain more accuracy prediction results than boosting and RS individually by rigorous theoretical analysis and experimental analysis. The pseudo code for the RS–boosting algorithm is given in Fig. 5 [8].

2.6 Multi-boosting

Multi-boosting is an extension to the highly successful AdaBoost technique for forming decision committees, which can be viewed as combing AdaBoost with wagging. Webb [16] considers that multi-boosting is advantageous to get more accuracy than AdaBoost and wagging individually. In this paper, we use C4.5 as the base learning algorithm of multi-boosting. Significantly, the wagging method actually is a paradigm of bagging method. Bauer and Kohavi [23] consider that wagging seeks to repeatedly perturb the training set as in bagging, but instead of sampling from it, wagging adds Gaussian noise to each weight with mean zero and a given standard deviation. Moreover, wagging is good at trading off bias and variance. The pseudocode for the multi-boosting algorithm is given in Fig. 6 [16].

3 Empirical study

Since the practices of SCF started only a couple of years ago and there are still many perplexing issues haunting the decision makers of SCF, only a few China's SMEs and CEs cooperate in this financing pattern. It is impossible to gather adequate research data of SCF from the literature. Meanwhile, it is also difficult to collect primary data from SMEs, CEs, financial institutions, or logistics companies by interview and survey. In order to predict the credit risk of China's SMEs in SCF, the financial and non-financial data of selected listed SMEs and CEs are selected from database. Significantly, there is real trading relationship among these SMEs and CEs. In this paper, we assume that if these SMEs are short of capital and starve for financing, then they cooperate with CEs and financial institutions to apply the SCF.

3.1 Data preparation

Our data are basically collected and mined from the China Stock Market and Accounting Research (CSMAR) solution database (http://www.gtarsc.com). The data include the quarterly financial and nonfinancial data of 48 listed SMEs from Small and Medium Enterprise Board of Shenzhen Stock Exchange, six listed CEs from Shanghai Stock Exchange and three listed CEs from Shenzhen Stock Exchange during the period of 31 March 2012–31 December 2013. After deleting data points of unavailable entries, we retain 377 valid quarterly data points that can be are used for constructing SMEs credit risk prediction model. The nine CEs have abundant financial resources and a high degree of credibility. The 48 listed SMEs include 12 star special treatment (*ST) listed companies and 36 non-star special treatment (non-*ST) listed companies. The *ST listed SMEs are defined as the listed companies in Small and Medium Enterprise Board of Shenzhen Stock Exchange, that suffer operating losses for two consecutive years and face with delisting warning. In this study, each quarterly data sample of *ST SMEs release a ‘negative signal’ in the two years before they are labeled *ST; on the contrary, each quarterly data sample of non-*ST SMEs release a ‘positive signal’ in the past consecutive two years. Accordingly, we categorize 48 SMEs into two groups, the extremely high credit risk group and the relatively low credit risk group according to whether the SME is a *ST listed company.

In this study, 18 financial and non-financial variables are chosen as the criteria to measure the SMEs credit risk in SCF, according to the suggestion of Xiong et al. [26]. These 18 variables serve as the independent variables of six classifier models in this paper, which cover applicant factor, counter party factor, items’ characteristics factor and operation condition factor (see Table 1). The dependent variable represents whether each quarterly data sample of SME release the high credit risk signal: value 0 means a ‘negative signal’ which signifies that the SME’s compliance probability is low and credit risk is high, while value 1 means a ‘positive signal’ signal which signifies that the SME’s compliance probability is high and credit risk is low.

Table 1 Financial and non-financial variables for predicting SMEs credit risk in SCF

Factors	Code	Variables
Applicant factors	R1	Current ratio
	R2	Quick ratio
	R3	Cash ratio
	R4	Working capital turnover
	R5	Return on equity
	R6	Profit margin on sales
	R7	Rate of return on total assets
	R8	Total assets growth rate
Counter party factors	R9	Credit rating of CEs
	R10	Quick ratio
	R11	Turnover of total capital
	R12	Profit margin on sales
Items’ characteristics factors	R13	Price rigidity, liquidation and vulnerable degree of trade goods
	R14	Account receivable collection period
	R15	Accounts receivable turnover ratio
Operation condition factors	R16	Industry trends
	R17	Transaction time and transaction frequency
	R18	Credit rating of SMEs

3.2 Evaluation criteria

The experiments’ evaluation criteria are adopted from the established standard measures in the fields of predicting SMEs credit risk of SCF. These measures include average accuracy, type I error, type II error, ‘precision’ rate, ‘recall’ rate, ‘F-Measure’ rate and receiver operating characteristic (ROC) curve.

The average accuracy includes ‘correctly classified instances’ and ‘incorrectly classified instances’. The former means that the absolute number and percentage of correctly classified instances, and the latter measures means that the absolute number and percentage of incorrectly classified instances. The average accuracy is defined as

$$\text{Average accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}, \tag{1}$$

where FN, TP, FP, and TN denote the ‘false negative’, ‘true positive’, ‘false positive’, and ‘true negative’ respectively; ‘negative’ means risk and ‘positive’ means non-risk.

Type I error and Type II error are also known as ‘false negative rate’ and ‘false positive rate’. These two kinds of evaluation criteria are widely used for measuring the binary classification results, such as measuring the performance of corporate credit risk prediction methods [8, 27]. Wang and Ma [8] consider that classifier acquires a good prediction result because of reducing type II error. Type I error and Type II error are, respectively, defined as

$$\text{Type I error} = \frac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}}, \tag{2}$$

$$\text{Type II error} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}. \tag{3}$$

Meanwhile, we apply other important evaluation criteria, i.e., ‘F-Measure’ rate as suggested by Power [27]. The ‘F-Measure’ rate is defined as Eq. (4). ‘F-Measure’ rate is also called ‘F_1’ rate which is the arithmetic mean of ‘precision’ rate and ‘recall’ rate [27]. The ‘precision’ rate and ‘recall’ rate are defined as Eqs. (5) and (6). The ‘precision’ rate (also called positive predictive value) denotes the proportion of ‘Predicted Positive’ cases that are correctly ‘Real Positives’, while the ‘recall’ rate (also known as sensitivity) is the proportion of ‘Real Positive’ cases that are correctly ‘Predicted Positive’ [27]. The higher the ‘precision’ rate is, the lower the 'false positive rate' of classifier is. A high 'recall' rate corresponds to a high 'true positive rate' of classifier. Equation (4) shows that a high value of 'F_1' rate can ensure a high value of 'precision' rate and 'recall' rate. Instead, the higher value of 'F_1' means that the classifier has a better prediction performance. Mathematically, the 'F-measure', the 'precision' rate, and the 'recall' rate are, respectively, defined as

$$F_1 = \frac{2}{\frac{1}{r} + \frac{1}{p}}, \tag{4}$$

$$p = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{5}$$

$$r = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{6}$$

where p and r mean 'precision' rate and 'recall' rate respectively.

Hosmer et al. [28] argue that a better and more complete description of classification accuracy is the area under receiver operating characteristic (ROC) curve. This curve, originating from signal detection theory, shows how the receiver detects the existence of signal in the presence of noise [28]. It plots the probability of detecting true signal (sensitivity) and false signal (1-specificity) for an entire range of possible cutpoints [28]. Hosmer et al. [28] give the general guidelines as follows: when $\mathrm{ROC} = 0.5$, this suggests no discrimination; when $0.5 < \mathrm{ROC} < 0.7$, this suggests poor discrimination; when $0.7 < \mathrm{ROC} < 0.8$, this suggests acceptable discrimination; when $0.8 < \mathrm{ROC} < 0.9$, this suggests excellent discrimination; when $\mathrm{ROC} \geq 0.9$, this suggests outstanding discrimination. As above definitions, the more value of area under ROC curve is adjacent to 1 the more classifier method is outstanding.

3.3 Experimental procedure

The experiments are performed on a PC with a 3.19 GHz Intel Core i3 CPU and 1.92 GB RAM, using Windows XP operating system. Data mining toolkit Waikato Environment for Knowledge Analysis (WEKA) version 3.6.12 is sued for experiment. WEKA is a popular and free available suite of machine learning and data mining software which is written with Java and developed at the University of Waikato, New Zealand.

We firstly compare average accuracies of IEML methods (i.e., multi-boosting and RS–boosting) with other three EML methods (i.e., bagging, boosting and RS) and an IML method (i.e., DT), for predicting SMEs' credit risk in SCF. Secondly, we compare type I and II errors of IEML methods with that of other three EML methods and an IML method. Finally, we compare area under ROC curves of IEML methods (i.e., multi-boosting and RS–boosting) with that of their base EML methods (i.e., bagging, boosting and RS) and IML (i.e., DT) method, respectively. For implementation of bagging, boosting, RS and DT, we choose WEKA bagging module, i.e., WEKA ADBoostM1 module, WEKA random subspace module and WEKA J48, respectively. For implementation of RS–boosting, we use WEKA Package, i.e., WEKA.JAR and implement in Eclipse according to Wang and Ma [8]. For implementation of multi-boosting, we use WEKA MultiBoostAB. Meanwhile, we employ DT as base classifier of multi-boosting and RS–boosting according to Maclin and Opitze [20], Fu et al. [21] and Wang and Ma [8].

In order to minimize the influence of the variability of the training set, ten times tenfold cross-validation is performed. In other words, the dataset is randomly divided into ten groups with similar sizes and distributions. The nine groups' subsets are used as training set while the remaining subset is used as the test set, which is repeated for ten times such that every subset is used as the test set once. The average of the ten groups' test results is taken as the final prediction accuracy rate of the model.

4 Empirical results

In this section, we firstly evaluate two IEML methods on 377 data sets using both NN (i.e., Multilayer Perceptron) and DT (i.e., C4.5) as the base classifier for choosing the appropriate one. Then, we show that IEML methods compete quite outstanding against EML methods and IML method by analyzing prediction evaluation criteria (i.e., average accuracy, type I error, II error, 'precision' rate, 'recall' rate, 'F-Measure' rate and ROC curve), which are excellent methods for predicting SMEs credit risk in SCF. Meanwhile, we compare the prediction evaluation criteria of multi-boosting with RS–boosting in order to find the better IEML method in predicting SMEs' credit risk.

4.1 The base classifier

As Maclin and Opitz [20] suggested, we give the explanation of using DT (i.e., C4.5) as the base classifier of

Table 2 Average accuracy results of two base classifiers

Methods	RS–Boosting	Multi-Boosting
Decision tree	85.41 %[a] (14.59 %)[b]	84.08 %[a] (15.92 %)[b]
Neural networks	82.76 %[a] (17.24 %)[b]	80.11 %[a] (19.89 %)[b]

[a] Correctly classified instances

[b] Incorrectly classified instances

IEML methods instead of using NN (i.e., Multilayer Perceptron) by comparing their average accuracies. Table 2 shows that both RS–boosting and multi-boosting get the better average accuracy when the DT is used as the base classifier. Therefore, DT is used as the base classifier of IEML methods in this study.

Table 3 Average accuracy results of six methods

Methods	Correctly classified instances (%)	Incorrectly classified instances (%)
DT	79.58	20.42
Bagging	77.19	22.81
Boosting	74.80	25.20
RS	77.45	22.55
RS–boosting	85.41	14.59
Multi-boosting	84.08	15.92

4.2 Prediction evaluation

Table 3 shows that RS–boosting has the highest average accuracy of 85.41 %. Closely following RS–boosting is another IEML method, i.e., multi-boosting which reaches average accuracy of 84.08 %. It is noteworthy that boosting obtains the lowest average accuracy among four methods, i.e., DT with 79.58 %, bagging with 77.19 %, RS with 77.45 %, and boosting with 74.80 %. It is interesting that the average accuracy result of boosting not only is the worst one among three EML methods, but also is worse than the IML (i.e., DT). In other words, the prediction accuracy of EML method is not always better than IML.

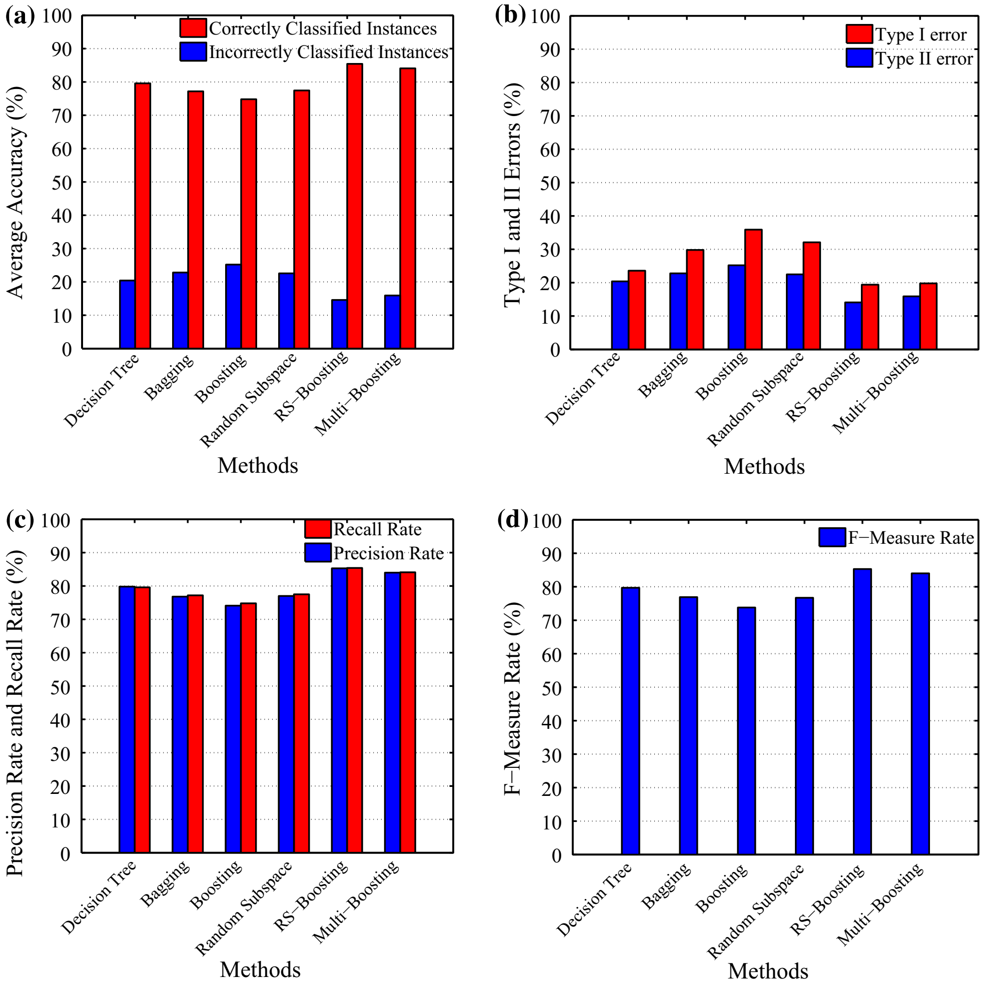

Fig. 7 Comparing the SMEs credit risk prediction accuracies of six methods. **a** The average accuracy of two integrated ensemble machine learning methods are both higher than that of other four methods, **b** the type I and II errors of two integrated ensemble machine learning methods are both lower than that of other four methods; **c** the 'precision' rate and 'recall' rate of two integrated ensemble machine learning methods are simultaneous higher than that of other four methods; **d** the 'F-Measure' rate of two integrated ensemble machine learning methods are simultaneous higher than that of other four methods (color figure online)

The poor performance of boosting results from over-fitting the training dataset since later training dataset may be over-emphasizing instances that are noise [20]. In order to visually compare the average accuracy results of six methods in details of correctly classified instances and incorrectly classified instances, we illustrate them in Fig. 7a.

From another point of view, the IEML methods of RS–boosting and multi-boosting acquire better results mainly depending on reducing type I and II errors. We present that RS–boosting gets the lowest type I and II errors of 19.40 and 14.10 % in Table 4. Closely following RS–boosting is multi-boosting with type I and II errors of 19.80 and 15.90 %. Thus, it also proves that RS–boosting is little better than multi-boosting. Significantly, according to West [13], Yap et al. [29], Kürüm et al. [30] and Bekhet and Eletter [31], we consider that the improvement in 'negative signal' prediction accuracy is more important than that of the 'positive signal' prediction accuracy for financial institution specially at the present stage of China's credit market. The type II error is the false positive rate, and instead, the method falsely classifies the 'negative signal' into 'positive signal'. Thus, we consider that reducing type II error is more important than reducing type I error. For that matter, we again prove that RS–boosting is the best prediction method among six methods. Moreover, we illustrate the type I and II errors of six methods in Fig. 7b.

In addition, we also research the prediction performance of classifiers base on 'precision' rate, 'recall' rate and 'F-Measure' rate which are suggested by Powers [27] and Alexander and Daniel [32]. Table 5 shows that RS–boosting gets the higher 'precision' rate, 'recall' rate and 'F-Measure' rate of 85.30, 85.40 and 85.30 %. Closely following RS–Boosting is Multi-Boosting with 'precision' rate, 'recall' rate and 'F-Measure' rate of 84.00, 84.10 and 84.00 %. Thus, it also proves that RS–boosting is little better than multi-boosting. Moreover, we illustrate the 'precision' rate, 'recall' rate and 'F-Measure' rate of six methods in Fig. 7c, d.

Subsequently, Table 6. shows that RS–boosting and multi-boosting methods demonstrate outstanding performance of discrimination. We further prove that IEML methods can enhance the performance of SMEs credit risk prediction. Meanwhile, the values of area under ROC curve show that RS–boosting is slighter outstanding than multi-boosting. In order to visually compare the areas under ROC curve of IEML methods with EML methods and DT method, we illustrate the ROC curves of DT, bagging, boosting and multi-boosting in Fig. 8a and illustrate the ROC curves of DT, RS, boosting and RS–boosting in Fig. 8b.

Table 4 Type I and II errors of six methods

Methods	Type I error (%)	Type II error (%)
DT	23.60	20.40
Bagging	29.80	22.80
Boosting	35.90	25.20
RS	32.10	22.50
RS–boosting	19.40	14.10
Multi-boosting	19.80	15.90

Table 5 'Precision' rate, 'recall' rate and 'F-Measure' rate of six methods

Methods	'Precision' rate (%)	'Recall' rate (%)	'F-Measure' rate (%)
DT	79.80	79.60	79.70
Bagging	76.80	77.20	76.90
Boosting	74.10	74.80	73.80
RS	77.00	77.50	76.70
RS–boosting	85.30	85.40	85.30
Multi-boosting	84.00	84.10	84.00

Table 6 Discrimination accuracies of six methods

Methods	Area under ROC curve	Discrimination accuracy
DT	0.860	Excellent
Bagging	0.846	Excellent
Boosting	0.813	Excellent
RS	0.858	Excellent
RS–boosting	0.910	Outstanding
Multi-boosting	0.907	Outstanding

5 Conclusion

In recent years, SCF becomes one of the primary ways for SMEs to obtain loans with looser constraints, for CEs to improve cash flow and for financial institutions to reduce possible risks. Meanwhile, prediction models of SMEs credit risk in supply chain finance (SCF) are particularly important for financial institutions to assess SMEs credit risk, improve supply chain cash flow, reduce possible risks of overall supply chain and make correct credit loan decisions. As for SMEs credit risk prediction in SCF, a slight improvement in accuracy might be translated into significant future saving. In this paper, we focus on increasing the accuracy of SMEs credit risk prediction due to a fraction of improvement in accuracy may further

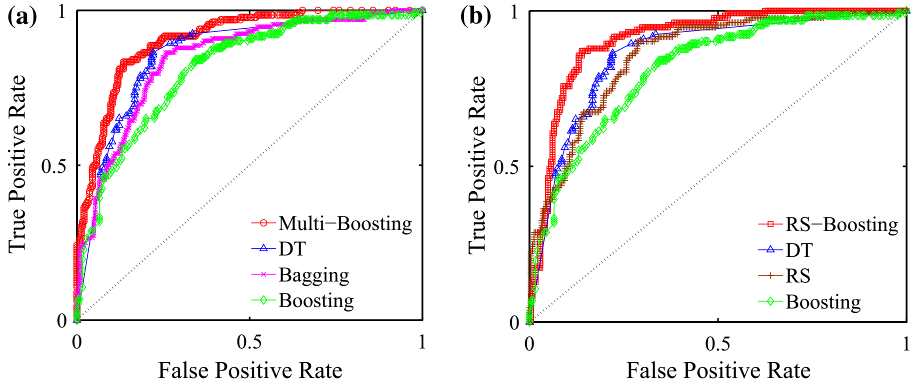

Fig. 8 ROC curves of six methods. Because RS–boosting method is integrated by boosting and RS, and multi-boosting method is integrated by boosting and wagging, meanwhile, DT is the base classifier of these two IEML methods, we compare areas under ROC curves of IEML methods (i.e., multi-boosting and RS–boosting) with that of their base EML methods (i.e., bagging, boosting and RS) and IML (i.e., DT) method, respectively in **a** and **b**. **a** The area of ROC curve of multi-boosting is enclosed by *red circles line*, which is obvious larger than the area of ROC curve of other methods; **b** the area of ROC curve of RS–boosting is enclosed by *red squares line*, which is obvious larger than the area of ROC curve other methods (color figure online)

guarantee normal operation of SMEs, CEs and financial institution in SCF.

A lot of prediction models of corporate credit risk are constructed based on the traditional statistical methods or IML methods such as logistic regression analysis and DT. Traditional statistical methods have some advantages such as efficient and robust. Simultaneously, researchers generally consider that IML methods have strong credit risk prediction capability and do not require the knowledge of the underlying relationships between input and output variables. However, EML method is a kind of powerful machine learning paradigm which has better advantages than IML method to predict corporate credit risk. Meanwhile, different EML methods have respective characteristics, which lead researchers to propose IEML methods. It should be noted that we apply two types of IEML methods, i.e., RS–boosting and multi-boosting, to predict SMEs credit risk in SCF. RS–boosting method is integrated by boosting and RS, and multi-boosting method is integrated by boosting and wagging.

Experiments are based on quarterly financial and non-financial data of 48 listed SMEs and nine listed CEs in China securities market during the period of 2012–2013 show that two IEML methods obtain better performance than other four methods, i.e., DT, bagging, boosting and RS. Moreover, we consider that RS–boosting is better than multi-boosting, which exhibits outstanding ability of SMEs credit risk prediction in SCF.

Acknowledgments This work was supported by the National Natural Science Foundation of China under Grant Nos. 71373072 and 71501066; the China Scholarship Council under Grant No. 201506135022; Specialized Research Fund for the Doctoral Program of Higher Education under Grant No. 20130161110031; and Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Grant No. 71221001.

References

1. More D, Basu P (2013) Challenges of supply chain finance: a detailed study and a hierarchical model based on the experiences of an Indian firm. Bus Process Manag J 19(4):624–647
2. Fairchild A (2005) Intelligent matching: integrating efficiencies in the financial supply chain. Supply Chain Manag Int J 10(4):244–248
3. Knox A (2005) Electronic payment: the missing link in supply chain efficiency. J Financ Transform Mark Imperfection 14:16–18
4. Hofmann E, Belin O (eds) (2011) Supply chain finance solutions. Springer, Berlin
5. Seifert RW, Seifert D (2011) Financing the chain. Int Commer Rev 10(1):32–44
6. Sopranzetti BJ (1999) Selling accounts receivable and the underinvestment problem. Q Rev Econ Finance 39(2):291–301
7. Wuttke DA, Blome C, Henke M (2013) Focusing the financial flow of supply chains: an empirical investigation of financial supply chain management. Int J Prod Econ 145(2):773–789
8. Wang G, Ma J (2011) Study of corporate credit risk prediction based on integrating boosting and random subspace. Expert Syst Appl 38(11):13871–13878
9. Huang Z, Chen H, Hsu CJ, Chen WH, Wu S (2004) Credit rating analysis with support vector machines and neural networks: a market comparative study. Decis Support Syst 37(4):543–558
10. Brill J (1998) The importance of credit scoring models in improving cash flow and collections. Bus Credit 100(1):16–17
11. Thomas LC (2000) A survey of credit and behavioral scoring: forecasting financial risks of lending to customers. Int J Forecast 16(2):149–172
12. Jiang Y (2009) Credit scoring model based on the decision tree and the simulated annealing algorithm. In: 2009 World congress on computer science and information engineering. Los Angeles, CA, pp 18–22
13. West D (2000) Neural network credit scoring models. Comput Oper Res 27(11):1131–1152

14. Nanni L, Lumini A (2009) An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. Expert Syst Appl 36(2):3028–3033
15. Tsai C, Wu J (2008) Using neural network ensembles for bankruptcy prediction and credit scoring. Expert Syst Appl 34(4):2639–2649
16. Webb GI (2000) Multiboosting: a technique for combining boosting and wagging. Mach Learn 40(2):159–196
17. Zhou Z (ed) (2009) Ensemble. In: Encyclopedia of database systems. Springer, New York
18. Hall MA, Smith LA (1998) Practical feature subset selection for machine learning. In: Computer science '98 proceedings of the 21st Australasian computer science conference ACSC '98, Berlin, pp 181–191
19. Quinlan JR (ed) (1993) C4.5: programs for machine learning. Morgan Kaufmann Publishers, San Francisco, CA
20. Maclin R, Opitz D (1999) Popular ensemble methods: an empirical study. J Artif Intell Res 11:169–198
21. Fu ZW, Golden BL, Lele S, Raghavan S, Wasil E (2006) Diversification for better classification trees. Comput Oper Res 33(11):3185–3202
22. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
23. Bauer E, Kohavi R (1999) An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach Learn 36(1–2):105–139
24. Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Thirteenth international conference on machine learning, San Francisco, pp 148–156
25. Ho TK (1998) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20(8):832–844
26. Xiong X, Ma J, Zhao W (2009) Credit risk analysis of supply chain finance. Nankai Bus Rev 12(4):92–98
27. Powers DMW (2011) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. J Mach Learn Technol 2(1):37–63
28. Hosmer JDW, Lemeshow S, Sturdivant RX (eds) (2013) Applied logistic regression. Wiley, Hoboken, NJ
29. Yap BW, Ong SH, Husain NHM (2011) Using data mining to improve assessment of credit worthiness via credit scoring models. Expert Syst Appl 38(10):13274–13283
30. Kürüm E, Yildirak K, Weber GW (2012) A classification problem of credit risk rating investigated and solved by optimization of ROC curve. Cent Eur J Oper Res 20(3):529–557
31. Bekhet HA, Eletter SFK (2014) Credit risk assessment model for Jordanian commercial banks: neural scoring approach. Rev Dev Finance 4(1):20–28
32. Alexander F, Daniel M (2007) Measuring word alignment quality for statistical machine translation. Comput Linguist 33(3):293–303