Expert Systems with Applications 39 (2012) 3668–3675

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Dynamic classifier ensemble model for customer classification with imbalanced class distribution

Jin Xiao [a,*], Ling Xie [b], Changzheng He [a], Xiaoyi Jiang [c]

[a] School of Business Administration, Sichuan University, Chengdu 610064, Sichuan Province, China
[b] School of Public Management, Sichuan University, Chengdu 610064, Sichuan Province, China
[c] University of Münster, Department of Mathematics and Computer Science, Einsteinstraße 62, 48149 Münster, Germany

ARTICLE INFO

Keywords:
Customer classification
Dynamic classifier ensemble
Cost-sensitive learning
Imbalanced class distribution

ABSTRACT

Customer classification is widely used in customer relationship management including churn prediction, credit scoring, cross-selling and so on. In customer classification, an important yet challenging problem is the imbalance of data distribution. In this paper, we combine ensemble learning with cost-sensitive learning, and propose a dynamic classifier ensemble method for imbalanced data (DCEID). For each test customer, it can adaptively select out the more appropriate one from the two kinds of dynamic ensemble approach: dynamic classifier selection (DCS) and dynamic ensemble selection (DES). Meanwhile, new cost-sensitive selection criteria for DCS and DES are constructed respectively to improve the classification ability for imbalanced data. We apply this method to a credit scoring dataset in UCI and a real churn prediction dataset from a telecommunication company. The experimental results show that the classification performance of DCEID is not only better than some static ensemble methods such as weighted random forests and improved balanced random forests, but also better than the existing DCS and DES strategies.

1. Introduction

Customer classification is an important issue in real world marketing. It aims at building a model to predict future customer behaviours through classifying database records into a number of predefined classes based on certain criteria (Ngai, Xiu, & Chau, 2009) and is widely used in customer churn prediction, credit scoring, cross-selling, and so on (Chan, 2008; Hwang, Jung, & Suh, 2004). The traditional customer classification methods such as decision tree (Luo & Mu, 2004), artificial neural network (ANN) (Yan, Miller, Mozer, & Wolniewicz, 2001), logistic regression (Kim & Yoon, 2004), genetic algorithm (Eiben, Koudijs, & Slisser, 1998) and support vector machine (SVM) (Coussement & Van den Poel, 2008; Huang, Chen, Hsu, Chen, & Wu, 2004) usually assume that the class distribution of the training data is balanced, that is to say the number of samples in each class is roughly the same. However, in real customer classification problems, the class distribution of customer data is often imbalanced. Therefore, the above algorithms cannot achieve satisfactory classification performance, which has posed a serious difficulty to customer classification (Xie, Li, Ngai, & Ying, 2009).

The imbalanced class distribution is characterized as having much more instances of some classes than others. Particularly for a bi-class application, the imbalance means that one class is represented by a large of samples, while the other one is represented by only a few. For example, in customer churn prediction, the number of churn customers is only a small proportion (usually 2% of the total samples) (Zhao & Dang, 2008); in customer credit scoring, customers with poor credit or fraudulent consuming are often only a small fraction in all customers (Fawcett & Provost, 1997). Traditional classification methods get poor performance in the class imbalanced data, mainly because they always generalize from the training set on the simplest hypothesis that best fits the data (Sun, Kamel, Wong, & Wang, 2007). This hypothesis seldom concerns about the minority sample of imbalanced dataset. Therefore, the classification rules that predict the minority class tend to be fewer and weaker than that of the majority class. Consequently, test samples belonging to the minority class are misclassified more often than those belonging to the majority class. However, in most application fields, the value of the correct classification for the minority customers is often greater than that for the majority ones. For instance, in customer churn prediction what we concern about most is whether the model can correctly predict the churn customers' behaviours. Thus, how to improve the classification ability of the algorithm for the class imbalanced data is one of the key issues to be solved in customer classification.

At present, there are two commonly used approaches to tackle the problem of class imbalance in customer classification. One is to use the resampling technique to balance the class distribution

* Corresponding author. Tel.: +86 28 85416236.
E-mail address: xjxiaojin@126.com (J. Xiao).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.09.059

first, and then construct classification model on the balanced training set (Laha, 2007; Liu, Hu, & Yu, 2008; Padmaja, Dhulipalla, Bapi, & Krishna, 2007). The most commonly used resampling methods include under-sampling for majority samples, over-sampling for minority samples, improved over-sampling method SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), as well as combination of under-sampling and over-sampling. For example, Laha (2007) has proposed a credit scoring model by integrating fuzzy rule based classification technique and *k*-nn method. To eliminate the impact of class imbalance on customer credit scoring, he adopts under-sampling to make the ratio of positive (minority) and negative (majority) customers in training set be 1:1. Padmaja et al. (2007) have proposed a new approach called extreme outlier elimination and hybrid sampling technique for fraud detection, in which they adopt hybrid sampling technique (a combination of SMOTE to over-sample the minority data and random under-sampling to under-sample the majority data) to improve the classification accuracy for imbalanced data.

The second approach is based on cost-sensitive learning. In customer classification with class imbalance, the misclassification cost of the customers belonging to different class is usually unequal. For instance, in customer churn prediction, the prediction that makes a churn customer as non-churn customer means the company is likely to lose the customer; while makes a non-churn customer as churn customer, the company will pay extra cost of customer retention. Practical experience shows that the cost to obtain a new customer is 4–6 times as large as to retain an old customer (Bhattacharya, 1998). Thus, this kind of approach assigns a higher cost to misclassification of the minority class, and tries to minimize the overall cost. There are many researches in this field (Bradford, Kunz, Kohavi, Brunk, & Brodley, 1998; Gama, 2000; Ting, 2002). Bradford et al. (1998) have studied the pruning methods for decision tree when the goal is minimizing the loss rather than the error, and found that applying the Laplace correction to estimate the probability distributions at the leaves is beneficial to all pruning methods. Gama (2000) has presented an iterative Bayes, which implements cost-sensitive learning by iterative updating. Ting (2002) has introduced an instance-weighting method to induce cost-sensitive trees, and the experimental results show that this algorithm is better than the standard decision tree.

On the other hand, the multiple classifiers ensemble (MCE) has been introduced into the customer classification of class imbalance in recent years. Lariviere and Van den Poel (2005) and Buckinx and Van den Poel (2005) have introduced the random forests to the customer churn prediction first. The comparison between random forests and some other methods such as neural network, regression analysis shows the advantages of ensemble learning algorithm in classification performance. Further, some scholars combine resampling technique with the ensemble learning for customer classification. For example, Zhou and Liu (2006) have combined the over-sampling technique with multiple classifiers combination, and proposed that such method can not only utilize the advantages of over-sampling technique to increase the number of minority samples so that the classifiers are able to improve the classification performance of minority samples better, but also utilize the advantages of multiple classifiers combination method to improve the overall classification performance of imbalanced dataset. Paleologo, Elisseeff, and Antonini (2010) have proposed a credit scoring model—Subagging based on ensemble learning. It selects all positive samples (assuming the positive samples are minority) and randomly takes a part of the negative samples without replacement, i.e., the under-sampling method to deal with the class imbalanced data. Finally, there are also some researches on combining cost-sensitive learning with ensemble learning. For instance, Chen, Liaw, and Breiman (2004) have presented weighted random forests (WRF). Xie et al. (2009) have proposed improved balanced random forests for customer churn prediction which combines WRF with the balanced random forests (Chen et al., 2004). Zhou, Lai, and Yu (2010) have proposed several least squares support vector machines ensemble models for credit scoring, in which cost-sensitive least support vector machine is regarded as the base classifier. Their experimental results show that ensemble strategies can improve the performance in some degree and are effective for building credit scoring models.

The propositions mentioned above have made important contributions to the customer classification with imbalanced class distribution. However, after careful analysis we can find that they all belong to the static classifier ensemble. In classifier ensemble, there are generally two kinds of approaches: static classifier ensemble (SCE) and dynamic classifier ensemble (DCE) (Kuncheva, 2002). SCE selects a unified ensemble scheme in the validation dataset for all test samples. However, in real customer classification problem, different test samples usually have different classification difficulty. Therefore, if we can select different classifier ensemble to classify according to the characteristics of each sample, the classification performance may be better than a simple SCE strategy, and this is just the basic idea of dynamic classifier ensemble (Ko, Sabourin, & Britto, 2008; Woods, Kegelmeyer, & Bowyer, 1997). At present, there are two kinds of commonly used dynamic classifier ensemble strategy: dynamic classifier selection (DCS) and dynamic ensemble selection (DES). The former is to select a single best classifier for each test sample, while the latter is to select an optimal classifier ensemble for each test sample. In our previous work (Xiao & He, 2009; Xiao, He, Jiang, & Liu, 2010), we have proposed GMDH (group method of data handling) based dynamic ensemble selection strategy (GDES), the experimental results show that the classification performance of GDES is statistically significant better than that of DCS. While we also have found that both DCS and DES have advantages and disadvantages. Given a test sample, if there is a classifier in the base classifier pool whose classification performance is significantly better than the others, DCS may be a better choice; while if there is no significant difference among the classification performance of all base classifiers, then DES may be better. Therefore, if we can combine these two strategies and make them complementary, the classification performance is expected to be improved further.

Therefore, this paper combines ensemble learning method with cost-sensitive learning, and proposes a dynamic classifier ensemble method for imbalanced data (DCEID). This method fuses DCS and DES effectively to improve the classification accuracy. For each test customer, DCEID can select out the more appropriate strategy from the DCS and DES, and adopt the strategy to classify the customer, thus achieve adaptive switching between the two strategies. At the same time, to improve the classification ability for imbalanced data, new cost-sensitive selection criteria are constructed for DCS and GDES in DCEID respectively. Further, we experimentally test DCEID and other seven ensemble strategies in a credit scoring dataset from UCI and a real customer churn prediction dataset from a telecommunication company.

The remainder of this paper is organized as follows. Section 2 introduces related work on multiple classifiers ensemble including some static ensemble methods and dynamic ensemble methods. Section 3 describes the DCEID strategy in detail. The experimental setup and detailed results is described in Section 4. Finally, some concluding remarks and ideas for future work are given in Section 5.

2. Related work on multiple classifiers ensemble

Many researchers have realized that there are many advantages for multiple classifiers ensemble (MCE) (Kittler, Hatef, Duin, & Matas, 1998; Dietterich, 2000a; Ghosh, 2002). For instance, Kittler

et al. (1998) have proposed two reasons for using MCE: efficiency and accuracy. It is observed that different classifiers do not necessarily misclassify the patterns simultaneously. Thus MCE can make the base classifiers complementary and improve the classification performance. Meanwhile, the combination of simple classifiers can usually achieve a similar accuracy as that of a more complicated classifier and significantly reduce the computation cost. These observations have motivated the relatively recent researches that adopt MCE for better classification performance (Chen & Kamel, 2009; Dos Santos, Sabourin, & Maupin, 2009; Goumas, Dimou, & Zervakis, 2010; Hsieh & Hung, 2010; Mallipeddi, Mallipeddi, & Suganthan, 2010; Paleologo et al., 2010; Twala, 2010; Zhou et al., 2010). In the past decades, various classifier ensemble strategies have been proposed and their usefulness has been experimentally demonstrated.

2.1. Static classifiers ensemble method

There are many static classifiers ensemble (SCE) strategies. Here we briefly introduce some commonly used ones.

2.1.1. Majority voting

Majority voting (MAJ) is the most common method to combine more than one decision. It can be divided into absolute majority voting and relative majority voting. The former means that only the class label which is the output of more than half basic classifiers can become the final output label. The latter means that the class label which gets more votes than any others will become the final class label. MAJ is very simple. While it has a disadvantage that all based classifiers are treated equally regardless of the characteristic of each classifier (Kim, Kim, & Lee, 2003). Thus, weighted majority voting (WMAJ) is an improved version of MAJ (Goldman & Warmuth, 1995). In WMAJ, each base classifier is assigned a weight according to its classification performance. For example, Sun and Li (2008) have applied WMAJ to listed companies' financial distress prediction and achieved satisfactory results.

2.1.2. Genetic algorithm based ensemble

In Genetic algorithm based ensemble (GAE), a genetic algorithm maps a problem onto a set of strings. Each string represents a potential solution, which must encode $K \times M$ real-valued parameters (where N is the number of the base classifier, and M is the number of sample class). Through genetic algorithm, we need to find the optima combinational coefficients (Zhou, Wu, & Tang, 2002).

2.1.3. Neural network based ensemble

In neural network based ensemble (NNE) a feed-forward neural network takes the outputs of the ensemble of classifiers along with the input feature vector to try and learn how to weight the different classifiers. These weights reflect the degree of confidence in each classifier. That is, it tries to understand and collect information that might be helpful in determining how to combine the different classification outputs to achieve a better performance (Lipnickas & Korbicz, 2004).

2.1.4. Random forests

Random forests (RF) (Breiman, 2001) are a combination of tree classifiers such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. RF adopts a random selection of features to split each node, and it has been demonstrated that RF can achieve very good classification performance and is robust against over-fitting. However, the standard random forests do not work well on datasets where data is extreme imbalance. Thus, Chen et al. (2004) have proposed two ways to handle the imbalanced data classification problem of random forests: weighted random forests (WRF) and balanced random forests (BRF). The former method is based on cost-sensitive learning, and the latter one is based on a sampling technique. Further, Xie et al. (2009) have proposed improved balanced random forests (IBRF) by combining WRF and BRF. There experimental results have shown that IBRF produces better prediction results than BRF and WRF.

2.2. Dynamic classifiers ensemble method

There are also many dynamic classifier ensemble strategies, and we introduce three kinds of strategies: dynamic classifier selection based on local accuracy (DCS-LA) (Woods et al., 1997), dynamic ensemble selection by K-nearest-oracles (Ko et al., 2008) and dynamic ensemble selection based on GMDH (Xiao & He, 2009).

2.2.1. Dynamic classifier selection based on local accuracy

Let a set C contain N classifiers $C_1, C_2, \ldots, C_N$, which have already been trained to solve the M-class classification task. For each test sample x^*, $R(x^*)$ is the local region made up of the K nearest neighbors of x^*. Let $LA_{j,K}(x^*)$ be the local accuracy estimate for classifier C_j. Then DCS-LA algorithm is defined as follows (Woods et al., 1997):

1. If all of the base classifiers C_j ($j = 1, 2, \ldots, N$) assign test sample x^* to the same class, then x^* will be assigned to this class; otherwise.
2. Compute $LA_{j,K}(x^*)$, $j = 1, 2, \ldots, N$.
3. Find the classifier C_i: $C_i | LA_{j,K}(x^*) = \max_j(LA_{j,K}(x^*))$, and assign x^* to the output of C_i.

We can see that the accuracy of DCS-LA strictly depends on the correctness of the LA estimate in step 2. Woods et al. (1997) have put forward two calculation methods: overall local accuracy (OLA) and local class accuracy (LCA). In addition, they point out in their study that, LCA has better classification performance than OLA in most cases.

2.2.2. Dynamic ensemble selection by K-nearest-oracles

K-nearest-oracles (KNORA) (Ko et al., 2008) regards the accuracy of ensemble as the selection criterion. For each test sample, KNORA simply finds its K nearest neighbors in the validation set first, and then selects an ensemble of classifiers which can correctly classify those neighbors in the validation set, before finally fusing all the classifiers in the ensemble by majority voting. Ko et al. (2008) have proposed four different strategies based on KNORA: KNORA-ELIMINATE (KN-E), KNORA-UNION (KN-U), KNORA-ELIMINATE-W (KN-E-W) and KNORA-UNION-W (KN-U-W), in which KN-E and KN-U are two simple voting strategies while KN-E-W and KN-U-W are weighted voting strategies.

2.2.3. Dynamic ensemble selection based on GMDH

We have introduced group method of data handling (GMDH) theory to DES, and presented a novel strategy GMDH-based dynamic ensemble selection (GDES) (Xiao & He, 2009). The general idea of the GDES algorithm is as follows: for each test sample x^*, it selects K nearest neighbors from the validation set to compose a local region of competence D_K, gets new candidate models by combining pairs of models of the previous layer from the initial model sets composed by the classification results in D_K through all base classifiers, estimates their parameters in the model learning set by least square (LS), calculates the external criterion values of middle candidate models in the model selecting set and selects some of the best models as the input of the next layer. Repeat the

above process until we get the optimal complexity classifier ensemble for x^*.

The GDES strategy regards the symmetric regularity criterion (SRC) in GMDH as the external criterion evaluating the middle candidate models, which can only measure the accuracy of the ensemble. Thus, we have improved the simple GDES and proposed GDES-AD strategy for noise data further (Xiao et al., 2010). This strategy takes into account both accuracy and diversity of ensembles in the process of ensemble selection. Experimental results show that GDES-AD has stronger noise-immunity ability than other strategies.

3. The proposed DCEID

3.1. Basic idea of DCEID

In customer classification, many problems can be seen as a binary classification issue. For example, in customer churn prediction, customers are divided into two classes: "churn" and "non-churn"; in customer credit scoring, customers are divided into "good credit" and "bad credit", and so on. Therefore, in this paper we regard the binary classification as an example of customer classification, which can be easily extended to multi-class classification.

Generally, the dynamic classifier ensemble (DCE) includes two kinds of strategies: dynamic classifier selection (DCS) and dynamic ensemble selection (DES). For the GDES strategy (Xiao & He, 2009), the experimental results show that the classification performance of GDES is statistically significant better than that of the existing dynamic classifier selection method DCS-LA. However, as stated by the "No Free Lunch" theorem (Corne & Knowles, 2003), no algorithm may be assumed to be better than any other algorithm when its performance is averaged over all possible classes of problems. Therefore, to improve the classification performance, the DCEID strategy in this paper can select an appropriate ensemble strategy from DCS-LA and GDES for each test customer, which achieves adaptive switching between the two dynamic classifier ensemble strategies. Further, DCEID combines the cost-sensitive learning method with the ensemble method. It assigns different misclassification cost to the customer in different class, utilizes the cost-sensitive technology to improve the external evaluation criteria of existing GDES, and then constructs cost-sensitive GDES strategy and cost-sensitive DCS-LA. The position of DCEID in the entire classifier ensemble system is in Fig. 1.

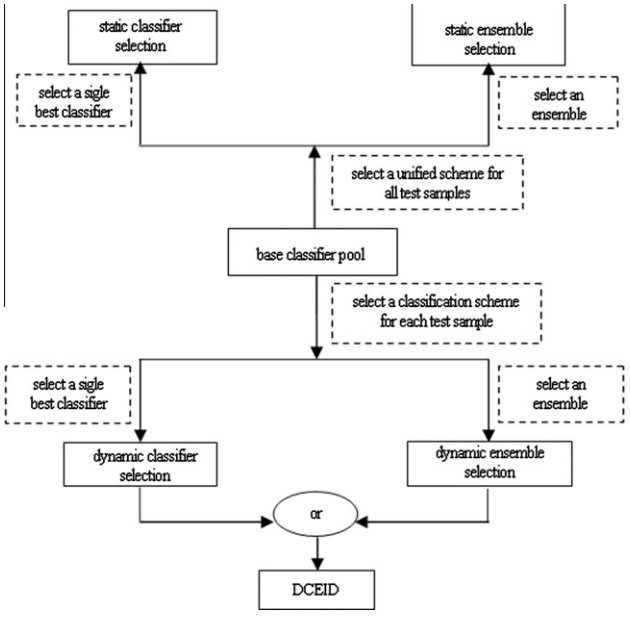

Fig. 1. The composition of multiple classifiers ensemble system.

3.2. The cost-sensitive evaluation criteria

In the GDES strategy proposed in our previous work (Xiao & He, 2009), we regard the symmetric regularity criterion (SRC) in GMDH as the external criteria of ensemble selection. For each test sample x^*, GDES strategy first find its K nearest neighbors from the validation set to compose a local competence region D_K, and then D_K is equally divided into two parts at random: subsets A and B. The SRC can be defined as follows:

$$d^2(W) = \Delta^2(A) + \Delta^2(B) = \sum_{t\in B}(y_B - y_B(A))^2 + \sum_{t\in A}(y_A - y_A(B))^2, \quad (1)$$

where y_B is the actual output of set B, $y_B(A)$ is the forecast output of set B by the model constructed in set A. Therefore, $\Delta^2(A)$ in Eq. (1) indicates the classification error in set B by the model constructed in set A, and $\Delta^2(B)$ indicates the classification error in set A by the model constructed in set B.

In SRC, all the test samples are assigned the same misclassification cost. However, the customers in different class usually have unequal misclassification cost in the customer classification problems with imbalanced class distribution. Thus, in this paper we combine cost-sensitive learning with SRC, and construct a new cost-sensitive symmetric regularity criterion (CS-SRC) for GDES,

$$S = Cost(A) + Cost(B), \quad (2)$$

where,

$$Cost(A) = \sum_{t=1}^{n_{11}} k(y_B - y_B(A))^2 + \sum_{t=1}^{n_{12}}(y_B - y_B(A))^2,$$

$$Cost(B) = \sum_{t=1}^{n_{21}} k(y_A - y_A(B))^2 + \sum_{t=1}^{n_{22}}(y_A - y_A(B))^2.$$

Here, we suppose that the misclassification cost of each majority sample equals 1, while k is the misclassification cost of each minority sample, i.e., the misclassification cost rate is 1: k; n_{11} is the number of minority samples and n_{12} is the number of majority samples in subset A; n_{21} is the number of minority samples and n_{22} is the number of majority samples in subset B. Therefore, $Cost(A)$ is the overall misclassification cost in set B by the model constructed in set A, $Cost(B)$ is the overall misclassification cost in set A by the model constructed in set B.

On the other hand, in standard dynamic classifier selection method DCS-LA (Woods et al., 1997), the classification error in the local competence region D_K (usually composed by K nearest neighbors of the test sample) is regarded as the criterion for evaluating the base classifiers, which can be described as follows:

$$\varepsilon_i = \sum_{t\in D_K}(y_t - y_t(C_i))^2, \quad (i = 1, 2, \ldots, N), \quad (3)$$

where y_t is the actual output of D_K, and $y_t(C_i)$ is the forecast output by the i-th classifier C_i. Then the standard DCS-LA selects the best single classifier with lowest classification error to classify the test sample.

In Eq. (3), all the samples in the local competence region are assigned the same misclassification cost, which is not suitable for the customer classification with imbalanced class distribution. Thus, in this paper we propose an improved cost-sensitive criterion for DCS-LA. Let D_{minor} and D_{major} be two mutually exclusive subsets in D_K: the minority class set and majority class set respectively, and then the improved criterion can be defined as follows:

$$\varepsilon_i' = \sum_{t \in D_{minor}} k(y_t - y(C_i))^2 + \sum_{t \in D_{major}} (y_t - y(C_i))^2, \quad (i = 1, 2, \ldots, N). \tag{4}$$

It can be seen that the former part of Eq. (4) is the misclassification cost in D_{minor} and the latter part is the misclassification cost in D_{major}.

3.3. Algorithm description

Let D_T, V, T_e be the training set, validation set and testing set of a customer classification problem with class imbalanced distribution respectively, and T_e contains m patterns. The basic steps of DCEID are as follows:

1. Train N base classifiers $C_1, C_2, \ldots, C_N$ in training set D_T as the base classifier pool.
2. For each test sample $x_j^* \in T_e$, $j = 1, 2, \ldots, m$,
   - 2.1 Find K nearest neighbors of x_j^* from set V to compose a new training set $D_K = (x_1, x_2, \ldots, x_K)$, i.e., a local competence region.
   - 2.2 Divide D_K into two subsets: minority class D_{minor} and majority class D_{major}, and assign the misclassification cost k and 1 to the test samples in D_{minor} and D_{major} respectively.
   - 2.3 Implement the cost-sensitive DCS-LA method. Calculate the overall misclassification cost of each base classifier C_i ($i = 1, 2, \ldots, N$) in local competence region D_K according to Eq. (4), find the best single classifier with the lowest misclassification cost, we call it C_{opt}.
   - 2.4 Conduct the cost-sensitive GDES strategy. Regard the CS-SRC in Eq. (2) as the external criterion of GDES and select the optimal classifier ensemble E_{opt} from the base classifier pool with the lowest external criteria value in D_K (for the detailed steps of GDES, please see the reference (Xiao et al., 2010)).
   - 2.5 Compare the misclassification cost of C_{opt} and E_{opt} in D_K (suppose they are ε_1 and ε_2), if $\varepsilon_1 < \varepsilon_2$, then classify x_j^* with the optimal single classifier C_{opt}; otherwise, classify x_j^* with the optimal ensemble E_{opt}.

4. Experimental analysis

4.1. Datasets and experimental setup

"German" is a famous UCI dataset (Merz & Murphy, 1995) on Germany credit scoring. It includes 20 attributes (7 numerical attributes and 13 qualitative attributes), and 1 class variable with two different states {*good*, *bad*} which divides the customers into two classes: customers with good credit and customers with bad credit. The dataset includes 1000 customer samples, 700 with good credit and 300 with bad credit. Thus, it belongs to the customer classification problem with imbalanced class distribution and its class imbalance rate *IR* (the rate between majority sample and minority sample) is 2.3333. It does not include missing data, meanwhile, it comes from UCI database and does not include noise (Dietterich, 2000b). To get training set, validation set and testing set, the whole dataset was divided into three parts equally at random, one third samples for training, one third for validation, and the remains for testing.

The other dataset is the customer churn prediction data of a branch of Sichuan Telecommunication ("telecom" dataset for short), and its interval is 2005.01–2005.06. According to the basic principle of attribute selecting for customer churn prediction and considering the availability of the data, we selected the following ten churn attributes: customer level, intra-regional call charges, range call charges, monthly fee, charges for domestic long-distance telephone, charges for international long-distance telephone, monthly total fee, average times out of service in three months, average expenditure in three months, total arrears. Further, for the customer class variable, we defined the off-network customer (churn customer) as those had charge record in this month while no next month. For example, an off-network customer in January 2008 is who has charge record in January and no in February. After simple data cleaning, we got 3350 samples from the database, among which 424 are churn customers, and 2926 are non-churn customers. The customer churn rate is 12.66% and its class distribution is highly imbalanced. We still divided the whole dataset into three parts equally at random, one third for training, one third for validation, and the remains for testing.

We compared the classification performance of DCEID strategy with some resampling based methods and cost-sensitive based methods. In resampling based methods, we first adopted the SMOTE algorithm (Chawla et al., 2002) to over-sample the minority data, then implement the static classifier ensemble methods: weighted majority voting (WMAJ) (Sun & Li, 2008), genetic algorithm based ensemble (GAE) (Zhou et al., 2002), neural network based ensemble (NNE) (Lipnickas & Korbicz, 2004), as well as the dynamic classifier ensemble methods: dynamic classifier selection based on local accuracy (DCS-LA) (Woods et al., 1997), KNORA-ELIMINATE (KN-E) (Ko et al., 2008). To be convenience, the five strategies above were denoted S-WMAJ, S-GAE, S-NNE, S-DCS and S-KNE. While in cost-sensitive based methods, weighted random forests (WRF) (Chen et al., 2004) and improved balanced random forests (IBRF) (Xie et al., 2009) were selected as the bench mark.

In WRF and IBRF, CART algorithm is used to generate the base decision tree classifier. Therefore, to make the classification results of different algorithms comparable, we also selected CART as the basic classification algorithm for all the multiple classifiers ensemble strategies compared in this paper. As for the size of base classifier pool, it is demonstrated that the dynamic classifier ensemble strategies give good results already for small ensembles of five classifiers and almost reach the highest classification performance when the pool size is around 25, while the accuracy of static ensemble approaches continues to increase slowly with the addition of new classifiers, and usually reaches its maximum with 50 or 100 classifiers (Tsymbal, Puuronen, & Patterson, 2003). Thus, we let the pool size of dynamic ensemble strategies including S-DCS, S-KNE and DCEID be 25, and for the static ensemble strategies including S-WMAJ, S-GAE, S-NNE, WRF and IBRF, the pool size was 100.

All the experiments were performed on the MATLAB 6.5 platform with a dual-processor 3.0 Ghz Pentium 4 Windows computer, and in each case the final classification result was the average of 10 experiments. As for the parameter K in DCEID strategy, we have experimented with seven different values of K: 3, 5, 7, 9, 10, 13 and 15, and found that the classification performance of GDES-ADs corresponding to $K = 7$ and $K = 10$ is significantly better than that of GDES-ADs corresponding to $K = 3$, $K = 15$, $K = 13$ and $K = 5$. And there is no statistically significant difference among the GDES-ADs corresponding to $K = 7$, $K = 9$ and $K = 10$ (Xiao et al., 2010). In this paper, we let $K = 10$ after repeated experiments. While for k, some scholar (Bhattacharya, 1998) has found that getting a new customer is 4–6 times the cost of keeping an old one. The cost of getting new customer is just a good approximation to the misclassification cost of minority customer, and the cost of keeping an old customer is also an approximation to the misclassification cost of majority customer. Therefore, we let $k = 5$ in this paper.

To evaluate the customer classification performance of various strategies under the condition of imbalanced class distribution, we introduced the evaluation matrix (see Table 1). Further, we introduced four commonly-used evaluation criteria:

Table 1
Evaluation matrix for two-dimensional customer classification.

State of sample	Predicted negative class	Predicted positive class	Total
Actual negative class (majority)	A	B	$A+B$
Actual positive class (minority)	C	D	$C+D$
Total	$A+C$	$B+D$	$A+B+C+D$

Note: A is the number of samples that are both actual negative class and predicted negative class; B is the number of samples that are actual negative class and predicted positive class; C is the number of samples that are actual positive class and predicted negative class; D is the number of samples that are actual positive class and predicted positive class.

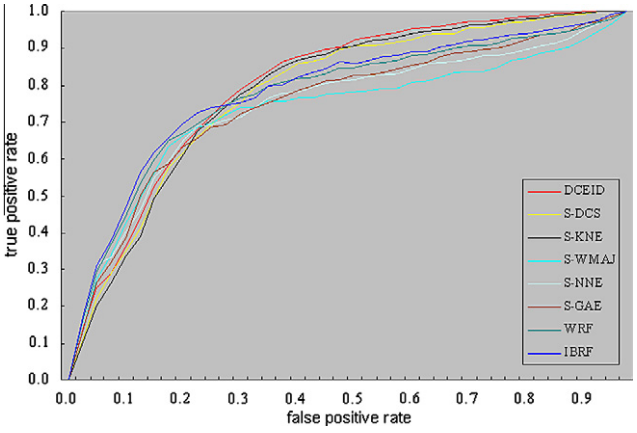

Fig. 2. ROC curves of eight strategies in "German" dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(1) total accuracy $= \frac{A+D}{A+B+C+D} \times 100\%$;
(2) minority accuracy $= \frac{D}{C+D} \times 100\%$;
(3) majority accuracy $= \frac{A}{A+B} \times 100\%$;
(4) ROC curve and AUC. The ROC curve for two classes problem is a figure about true positive rate – false positive rate, where the horizontal axis indicates the false positive rate $= B/(A+B) \times 100\%$ and the vertical axis indicates the truth positive rate $= D/(C+D) \times 100\%$. Because it is inconvenient to compare ROC curves of different classification strategies directly, people usually adopt the area under ROC curve (AUC) to compare the classification performance (Bradley, 1997).

4.2. Experimental results on "German" dataset

The classification results of eight strategies in "German" dataset are shown in Fig. 2 and Table 2. Fig. 2 displays the ROC curves of 8 strategies first, in which, the larger the area under the ROC curve is, the better the classification performance of the algorithm is. It can be seen from Fig. 2 that the ROC curve of DCEID strategy is higher than that of S-DCS and S-KNE, that is to say, the classification performance of DCEID is better than that of S-DCS and S-KNE in general. On the other hand, although the ROC curves of WRF, IBRF, S-WMAJ, S-GAE and S-NNE are higher than that of the DCEID in the former part, the area under the curve of each strategy is smaller than that of DCEID. Therefore, we can roughly conclude that the classification performance of DCEID is also better than that of WRF, IBRF, S-WMAJ, S-GAE and S-NNE in "German" dataset.

Further, the AUC, total accuracy, minority accuracy and majority accuracy of 8 strategies are shown in Table 2. It can be seen that the DCEID strategy has the highest AUC value, total accuracy and majority accuracy. Although the minority accuracy of DCEID is lower than that of S-KNE, we can still conclude that the DCEID strategy proposed in this paper has the best classification performance in the "German" dataset. It is worth noting that the former

Table 2
The classification performance comparison of eight strategies in "German" dataset.

Method	AUC	Total accuracy	Minority accuracy	Majority accuracy
DCEID	**0.8203**	**0.7893**	0.7574	**0.8030**
S-DCS	0.8132	0.7778	0.7548	0.7877
S-KNE	0.8176	0.7823	**0.7602**	0.7918
S-WMAJ	0.7599	0.7606	0.7184	0.7787
S-GAE	0.7956	0.7674	0.7560	0.7723
S-NNE	0.7823	0.7616	0.7234	0.7780
WRF	0.8024	0.7722	0.7362	0.7877
IBRF	0.8088	0.7757	0.7535	0.7852

Note: The boldface in each column indicates the maximum value of the corresponding evaluation criterion.

three strategies DCEID, S-DCS and S-KNE in Table 2 have higher AUC value, total accuracy, minority accuracy and majority accuracy than the other five strategies except that the minority accuracy of S-DCS is slightly worse than that of S-GAE. Therefore, the classification performance of DCEID, S-DCS and S-KNE is better than that of the other five strategies. This may be due to the reason that the former three strategies in Table 2 belong to dynamic classifier ensemble method, and the latter five strategies all belong to static classifier ensemble method. It has been demonstrated that dynamic classifier ensemble methods usually outperform the static classifier ensemble methods (Ko et al., 2008; Woods et al., 1997).

Finally, there are still some differences among the five static classifier ensemble methods. It can be seen from Table 2 that the AUC, total accuracy, minority accuracy and majority accuracy of WRF and IBRF are higher than that of S-WMAJ, S-GAE and S-NNE except that the minority accuracy of WRF is lower than that of S-GAE. In the five static ensemble methods, WRF and IBRF belong to cost-sensitive based method, and the other three strategies belong to resampling based method. Thus, we can conclude that the two cost-sensitive based methods have better classification performance than the three resampling based methods in "German" dataset.

4.3. Experimental results on "telecom" dataset

Fig. 3 displays the ROC curves of 8 strategies in "telecom" dataset. It can be seen that the ROC curve of DCEID strategy is higher than that of the other seven strategies. We can conclude roughly that the classification performance of DCEID is better than that of other seven strategies. Meanwhile, the ROC curves of 8 strategies assemble into three clusters approximately: the first cluster including DCEID, R-DCS and R-KNE belongs to dynamic classifier ensemble method; the second cluster including IBRF and WRF belongs to static ensemble method based on cost-sensitive learning; the third cluster including R-NNE, R-GAE and R-MAJ belong to static classifier ensemble method based on resampling technology. It can also be seen from Fig. 3 that the ROC curves of DCEID, R-DCS and R-KNE are higher than that of IBRF, WRF, R-NNE, R-GAE and R-MAJ, which indicates the classification performance of DCEID, R-DCS and R-KNE is better than that of the other five strategies in some degree.

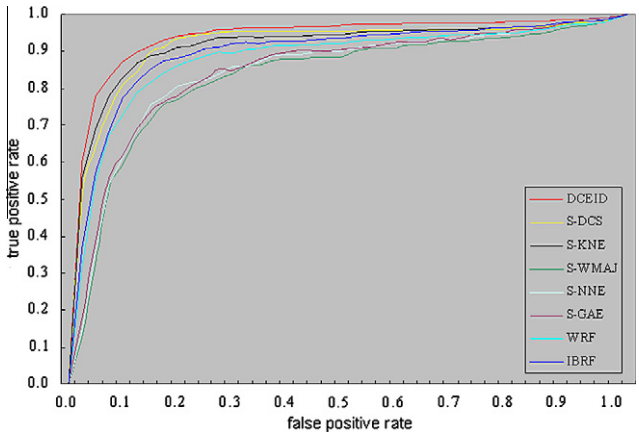

Fig. 3. ROC curves of eight strategies in "telecom" dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
The classification performance comparison of eight strategies in "telecom" dataset.

Method	AUC	Total accuracy	Minority accuracy	Majority accuracy
DCEID	**0.9683**	**0.9206**	**0.9181**	0.9206
S-DCS	0.9584	0.9184	0.9036	**0.9210**
S-KNE	0.9568	0.9172	0.9151	0.9175
S-WMAJ	0.9206	0.9069	0.9080	0.9067
S-GAE	0.9278	0.9115	0.9064	0.9123
S-NNE	0.9245	0.9052	0.9030	0.9055
WFF	0.9463	0.9167	0.9128	0.9173
IBRP	0.9532	0.9151	0.9145	0.9152

Note: The boldface in each column indicates the maximum value of the corresponding evaluation criterion.

In Table 3, the AUC, total accuracy, minority accuracy and majority accuracy of eight strategies in "telecom" dataset are presented. It can be seen that the DCEID strategy has the highest AUC value, total accuracy and minority accuracy. Although the majority accuracy of DCEID is slightly lower than that of S-DCS, we can still conclude that the DCEID strategy has the best classification performance. At the same time, DCEID, S-DCS and S-KNE have the higher AUC value, total accuracy, minority accuracy and majority accuracy than the other five strategies except that the minority accuracy of S-DCS is worse than that of IBRF and WRF. This demonstrates again that the dynamic classifier ensemble strategies usually can achieve better classification performance than static classifier ensemble strategies. Furthermore, it can also be seen that WRF and IBRF have higher AUC, total accuracy, minority accuracy and majority accuracy than S-NNE, S-GAE and S-WMAJ, which indicates that the two cost-sensitive learning based methods have better classification performance than the three resampling technology based methods in "telecom" dataset.

5. Conclusions and future work

In this paper, we mainly focus on the customer classification with imbalanced class distribution. It combines ensemble learning with cost-sensitive learning and proposes dynamic classifier ensemble method for imbalanced data (DCEID). On the one hand, the algorithm fuses two kinds of dynamic classifier ensemble approach: dynamic classifier selection (DCS) and dynamic ensemble selection (DES) effectively to improve customer classification accuracy. For each test customer, it can select a better strategy from dynamic classifier selection based on local accuracy (DCS-LA) and dynamic ensemble selection based on GMDH (GDES) for each test customer. Further, to improve the classification ability of DCEID for imbalanced data, we construct new cost-sensitive selection criteria for these two strategies. The experiments are conducted in a UCI customer credit scoring dataset and the customer churn prediction dataset of a real telecommunication company, and the results show that DCEID can manage the class imbalance problem in customer classification better and has better customer classification performance compared with some static ensemble strategies such as weighted random forests and improved balanced random forests, as well as the existing DCS and DES strategies.

In DCEID strategy, the cost-sensitive criterion in Eq. (2) only considers the misclassification cost in the process of ensemble selection, and it belongs to accuracy criterion. Actually, the diversity of ensembles is as important as the accuracy in ensemble selection. Therefore, if we can take into account the accuracy and diversity of the ensemble simultaneously in the process of selection, it may improve the classification performance further. In the future, we will explore how to effectively combine the cost-sensitive criterion with diversity measure of the ensemble for customer classification with imbalanced class distribution.

Acknowledgments

The authors thank the anonymous referees and the editor for their helpful comments. This research is supported by the Natural Science Foundation of China under Grant Nos. 71101100 and 71071101, Soft Science Program of Sichuan Province under No. 2010ZR0132, Research Start-up Project of Sichuan University under No. 2010SCU11012, and Sichuan University's Special Research Program for the Philosophy Social Science from the Subordinate Universities of Ministry of Education's Basic Research Foundation under Grant No. SKX201004.

References

Bhattacharya, C. B. (1998). When customers are members: customer retention in paid membership contexts. *Journal of the Academy of Marketing Science, 26*(1), 31–44.

Bradford, J., Kunz, C., Kohavi, R., Brunk, C., & Brodley, C. (1998). Pruning decision trees with misclassification costs. In *Proceedings of 10th European conference on machine learning* (pp. 131–136)). Berlin: Springer-Verlag.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research, 164*(1), 252–268.

Chan, C. C. H. (2008). Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. *Expert Systems with Applications, 34*(4), 2754–2762.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*(1), 321–357.

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data, Technical Report 666. Statistics Department of University of California at Berkeley.

Chen, L., & Kamel, M. S. (2009). A generalized adaptive ensemble generation and aggregation approach for multiple classifier systems. *Pattern Recognition, 42*(5), 629–644.

Corne, D. W., & Knowles, J. D. (2003). No free lunch and free leftovers theorems for multiobjective optimization problems. In *Evolutionary Multi-Criterion Optimization* (pp. 327–341). Berlin: Springer.

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications, 34*(1), 313–327.

Dietterich, T. G. (2000a). Ensemble methods in machine learning. *Lecture Notes in Computer Science, 1857*, 1–15.

Dietterich, T. G. (2000b). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning, 40*(2), 139–157.

Dos Santos, E. M., Sabourin, R., & Maupin, P. (2009). Overfitting cautious selection of classifier ensembles with genetic algorithms. *Information Fusion, 10*(2), 150–162.

Eiben, A. E., Koudijs, A. E., & Slisser, F. (1998). Genetic modelling of customer retention. *Lecture Notes in Computer Science, 1391*, 178–186.
Fawcett, T., & Provost, F. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery, 1*(3), 291–316.
Gama, J. (2000). A cost-sensitive iterative Bayes. In *Seventeenth international conference on machine learning, workshop on cost-sensitive learning* (pp. 7–13). San Francisco, CA: Morgan Kaufmann.
Ghosh, J. (2002). Multiclassifier systems: back to the future. *Lecture Notes in Computer Science, 2364*, 1–15.
Goldman, S. A., & Warmuth, M. K. (1995). Learning binary relations using weighted majority voting. *Machine Learning, 20*(3), 245–271.
Goumas, S. K., Dimou, I. N., & Zervakis, M. E. (2010). Combination of multiple classifiers for post-placement quality inspection of components: A comparative study. *Information Fusion, 11*(2), 149–162.
Hsieh, N. C., & Hung, L. P. (2010). A data driven ensemble classifier for credit scoring analysis. *Expert Systems with Applications, 37*(1), 534–545.
Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems, 37*(4), 543–558.
Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry. *Expert Systems with Applications, 26*(2), 181–188.
Kim, E., Kim, W., & Lee, Y. (2003). Combination of multiple classifiers for the customer's purchase behavior prediction. *Decision Support Systems, 34*(2), 167–175.
Kim, H. S., & Yoon, C. H. (2004). Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. *Telecommunications Policy, 28*(9), 751–765.
Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(3), 226–239.
Ko, A. H. R., Sabourin, R., & Britto, A. S. Jr., (2008). From dynamic classifier selection to dynamic ensemble selection. *Pattern Recognition, 41*(5), 1718–1731.
Kuncheva, L. I. (2002). Switching between selection and fusion in combining classifiers: An experiment. *IEEE Transaction on Systems, Man and Cybernetics – Part B, 32*(2), 146–156.
Laha, A. (2007). Building contextual classifiers by integrating fuzzy rule based classification technique and k-nn method for credit scoring. *Advanced Engineering Informatics, 21*(3), 281–291.
Lariviere, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications, 29*(2), 472–484.
Lipnickas, A., & Korbicz, J. (2004). Adaptive selection of neural networks for a committee decision. *International Scientific Journal of Computing, 3*(2), 23–30.
Liu, J., Hu, Q., & Yu, D. (2008). A comparative study on rough set based class imbalance learning. *Knowledge-Based Systems, 21*(8), 753–763.
Luo, N., & Mu, Z. C. (2004). Bayesian network classifier and its application in CRM. *Computer Application, 24*(3), 79–81.
Mallipeddi, R., Mallipeddi, S., & Suganthan, P. N. (2010). Ensemble strategies with adaptive evolutionary programming. *Information Sciences, 180*(9), 1571–1581.
Merz, C., & Murphy, P. (1995). UCI repository of machine learning databases. http://archive.ics.uci.edu/ml/.
Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications, 36*(2), 2592–2602.
Padmaja, T. M, Dhulipalla, N., Bapi, R. S, & Krishna, P. R (2007). Unbalanced data classification using extreme outlier elimination and sampling techniques for fraud detection. In *Proceeding of the 15th international conference on advanced computing and communications* (pp. 511–516). Washington DC: IEEE.
Paleologo, G., Elisseeff, A., & Antonini, G. (2010). Subagging for credit scoring models. *European Journal of Operational Research, 201*(2), 490–499.
Sun, J., & Li, H. (2008). Listed companies' financial distress prediction based on weighted majority voting combination of multiple classifiers. *Expert Systems with Applications, 35*(3), 818–827.
Sun, Y., Kamel, M. S., Wong, A. K. C., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition, 40*(12), 3358–3378.
Ting, K. M. (2002). An instance weighting method to induce cost-sensitive trees. *IEEE Transactions on Knowledge and Data Engineering, 14*(3), 659–665.
Tsymbal, A., Puuronen, S., & Patterson, D. W. (2003). Ensemble feature selection with the simple Bayesian classification. *Information Fusion, 4*(2), 87–100.
Twala, B. (2010). Multiple classifier application to credit risk assessment. *Expert Systems with Applications, 37*(4), 3326–3336.
Woods, K., Kegelmeyer, W. P., & Bowyer, K. (1997). Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(4), 405–410.
Xiao, J., & He, C. Z (2009). Dynamic classifier ensemble selection based on GMDH. In *Proceeding of the second international joint conference on computational sciences and optimization* (pp. 731–734). Washington DC: IEEE.
Xiao, J., He, C. Z., Jiang, X. Y., & Liu, D. H. (2010). A dynamic classifier ensemble selection approach for noise data. *Information Sciences, 180*(18), 3402–3421.
Xie, Y. Y., Li, X., Ngai, E. W. T., & Ying, W. Y. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications, 36*(3), 5445–5449.
Yan, L., Miller, D. J, Mozer, M. C, & Wolniewicz, R. (2001). Improving prediction of customer behavior in nonstationary environments. In *Proceeding of the international joint conference on neural networks* (pp. 2258–2263). Washington DC: IEEE.
Zhao, J., & Dang, X. H (2008). Bank customer churn prediction based on support vector machine: Taking a commercial bank's VIP customer churn as the example. In *4th International conference on wireless communications, networking and mobile computing* (pp. 1–4). Washington DC: IEEE.
Zhou, L., Lai, K. K., & Yu, L. (2010). Least squares support vector machines ensemble models for credit scoring. *Expert Systems with Applications, 37*(1), 127–133.
Zhou, Z. H., & Liu, X. Y. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering, 18*(1), 63–77.
Zhou, Z. H., Wu, J., & Tang, W. (2002). Ensembling neural networks: Many could be better than all. *Artificial Intelligence, 137*(1–2), 239–263.