Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2016, Article ID 5685970, 14 pages
http://dx.doi.org/10.1155/2016/5685970

Research Article

A Novel SMOTE-Based Classification Approach to Online Data Imbalance Problem

Chunlin Gong[1,2] and Liangxian Gu[1,2]

[1] *Northwestern Polytechnical University, Youyi West Road 127, Xi'an 710072, China*
[2] *National Aerospace Flight Dynamics Key Laboratry, Youyi West Road 127, Xi'an 710072, China*

Correspondence should be addressed to Chunlin Gong; leonwood@nwpu.edu.cn

Received 22 December 2015; Revised 10 April 2016; Accepted 14 April 2016

Academic Editor: Marek Lefik

In many practical engineering applications, data are usually collected in online pattern. However, if the classes of these data are severely imbalanced, the classification performance will be restricted. In this paper, a novel classification approach is proposed to solve the online data imbalance problem by integrating a fast and efficient learning algorithm, that is, Extreme Learning Machine (ELM), and a typical sampling strategy, that is, the synthetic minority oversampling technique (SMOTE). To reduce the severe imbalance, the granulation division for major-class samples is made according to the samples' distribution characteristic, and the original samples are replaced by the obtained granule core to prepare a balanced sample set. In online stage, we firstly make granulation division for minor-class and then conduct oversampling using SMOTE in the region around granule core and granule border. Therefore, the training sample set is gradually balanced and the online ELM model is dynamically updated. We also theoretically introduce fuzzy information entropy to prove that the proposed approach has the lower bound of model reliability after undersampling. Numerical experiments are conducted on two different kinds of datasets, and the results demonstrate that the proposed approach outperforms some state-of-the-art methods in terms of the generalization performance and numerical stability.

1. Introduction

With more and more successful real applications, machine learning acting as an efficient technique of data analysis and modeling is now becoming an important research field in the area of aeronautics science and mechanical engineering, for example, building surrogate model dynamically to support system design or identifying the characteristics of system by giving samplings. Among many concrete topics, online learning and imbalanced classification are both receiving a lot of attentions, and there also are many researches in the last decade for these two issues separately. However, to the best of our knowledge, there are few studies about the incorporation of these two topics, that is, data imbalance problem in online learning procedure. We also name it as online data imbalance problem which can be widely found in many real engineering applications such as fault diagnosis and damage detection. Therefore, studying this problem is of great significance.

In this paper, we try to provide an efficient solution from the perspectives of sampling strategy and learning algorithm. Considering the whole distribution characteristic of dataset and the feature of online learning, we present a novel SMOTE-based classification approach to the online data imbalance problem. This approach borrows the idea of granulation division to conduct oversampling and undersampling simultaneously. For major class, we use granule cores to replace the original samples for undersampling, while, for minor class, we firstly make granulation division and then conduct oversampling using SMOTE in the region around granule core and granule border. The above strategy is capable of following the data distribution of sample set, so it can be conducted repeatedly in the offline and online stages to balance the total sample set easily. Moreover, we introduce an efficient online learning algorithm, online sequential extreme learning machine (OS-ELM), which is combined with the proposed sampling strategy to achieve the fast and robust online learning for imbalanced data. To testify the effectiveness of the proposed approach, we first prove theoretically by means of the fuzzy information entropy that the proposed approach has the lower bound of model reliability after undersampling. And the experimental results on two different kinds of datasets also demonstrate the comparative performance of the proposed approach.

2. Related Works

Nowadays, the researches for the traditional data imbalance problem focus on two strategies [1]. One is data-based strategy which aims to make the original dataset balanced using mainly undersampling or oversampling techniques. The key idea of this kind of method is how to explore and obey the inner distribution characteristic of sample set in the sampling procedure. The synthetic minority oversampling technique (SMOTE) [2] is a widely used technique due to its simple form and straight idea, but it also suffers from the relatively low accuracy because SMOTE is incapable of exploiting the data distribution of sample set, especially in online case. Therefore, many SMOTE-based methods were developed with integrating other techniques. To improve the quality of synthetic samples, Verbiest et al. [3] introduced fuzzy-rough selection algorithm to reduce the noise generated by SMOTE after the balance stage. Gao et al. [4] utilized particle swarm optimization to optimize the undersampling procedure of SMOTE and then introduced RBF classification to reduce the misclassified cases. To reduce the imbalance level between classes, Zeng et al. [5] integrated the kernel trick and SMOTE into a new support vector machine (SVM) algorithm for data imbalance problem. As a combination with learning algorithm, Jeatrakul et al. [6] introduced SMOTE to neural networks in order to improve the generalization performance. Granular computing, known as an abstract idea for data processing, has instinct capacity to effectively remodel the original data according to data distribution, and then a high value of keeping the raw information of sample set is put. Therefore, granular computing has also been introduced to solve the data imbalance problem in SVM. For example, Wang et al. [7] utilized granulation division to hierarchically suppress the samples before SVM training. It is worth noticing that although the methods discussed above can balance sample set to some extent and thus improve the classification accuracy, they easily cause severe information loss if the distribution characteristic and feature are not considered well.

The other strategy is algorithm-based strategy which tries to improve the classification efficiency by developing the algorithms structure. For example, Hwang et al. [8] added weight factors to Lagrange multiplayer to improve the effectiveness of SVM upon facing imbalanced data. To lessen the misclassification rate, Yu et al. [9] calculated the moving distance of hyperplane by adjusting the decision threshold of SVM. Many other algorithms such as price-sensitive learning [10], weighted support vector machine [11], and weighted boost learning [12] were devoted to solve data imbalance problem. Although this strategy has been researched thoroughly, most of the algorithms can not apply to the online case directly due to lack of online structure. Besides, upon facing a large amount of data, it is generally hard for these algorithms to get results quickly. As extension form of single-hidden layer feedforward neural network (SLFN), extreme learning machines (ELMs), introduced by Huang et al. [13], have been recognized by their high learning speed and good generalization capacity for solving many problems of regression estimate and pattern recognition. As a sequential extension of ELM, online sequential ELM (OS-ELM) proposed by Liang et al. [14] can learn data one by one or chunk by chunk with fixed varied chunk size at very high speed. Although ELM has also been developed for data imbalance problem [15], it seems that OS-ELM has not been widely applied to data imbalance problems.

According to our literature survey, there are not too many researches about online data imbalance problem. By introducing prior duplication strategy, Vong et al. [16] firstly generated synthetic minority class samples and then utilized OS-ELM to establish an online sequential prediction model. Focusing on the modeling of data distribution in online pattern, Mao et al. [17] introduced the principal curve to exploit the inner structure of online data and then applied SMOTE to conduct oversampling by means of the distance from sample to principal curve. However, although this method could overcome many shortages of traditional methods, the principal curve is not well applicable to tackle the dataset with no apparent distribution feature. We noticed another recent work [18] for this problem which tried to adopt granulation division to remodel the distribution characteristic with a theoretical analysis about concrete information loss. Although it neglects some potential shortage of synthetic samples and the theoretical analysis needs to be improved largely, it is still an interesting attempt at this problem with reference value.

3. Background

3.1. SMOTE. SMOTE (synthetic minority oversampling technique) is a common oversampling method proposed by Chawla et al. [2]. In the SMOTE, instead of mere data oriented duplicating, the minority class is oversampled by creating synthetic instances in the feature space formed by the instance and its K-nearest neighbors, which effectively avoid the overfitting problem.

This method is described as follows. Choose two samples, x_1 and x_2, from the given minority sample set randomly, where each sample has n attributes. For x_1 and x_2, calculate the difference on the ith attribute; that is, $\text{diff}_i = x_{2i} - x_{1i}$. Then, we obtain the ith attribute value of the new target sample according to

$$x_{12i} = \text{rand}\,[0,1] * \text{diff}_i, \tag{1}$$

where rand$[0,1]$ means a random number between 0 and 1. So the final synthetic sample of x_1 and x_2 is

$$x_{12} = \text{rand}\,[0,1] * \text{diff}, \tag{2}$$

where $\text{diff} = (\text{diff}_1, \text{diff}_2, \ldots, \text{diff}_n)$.

According to the sampling rate we set execution times and repeat the above process. Incorporating the synthetic samples and the original samples, the final minority sample set is obtained.

3.2. Review of ELM and OS-ELM. As originally proposed for solving the single-hidden layer feedforward neural network

(SLFN), it has been proved that, with at most N hidden neurons, ELM can learn N distinct samples with zero errors by adopting any bounded nonlinear activation function [19]. Then, based on this approximation ability, ELM received wide attentions and has been developed into various forms, for example, multioutput regression [20]. The most important feature of ELM is its fast speed, owing to its single-hidden layer structure requiring no iterative process. In ELM, all the hidden node parameters are randomly generated without tuning. As an extension version of ELM, online sequential extreme learning machine (OS-ELM) is a faster and more accurate algorithm, which has been widely used in many fields, such as pattern recognition and data mining. The process of OS-ELM is divided into two steps: initialization phase and sequential learning phase and the detailed algorithm is described as follows [14].

Step 1 (initialization phase). Choose a small chunk $M_0 = \{(x_i, t_i), i = 1, 2, \ldots, N_0\}$ of initial training data, where $N_0 \geq \widetilde{N}$. Consider the following:

(1) Randomly generate the input weight $\mathbf{w}_i$ and bias b_i, $i = 1, 2, \ldots, \widetilde{N}$. Calculate the initial hidden layer output matrix $\mathbf{H}_0$.

(2) Calculate the output weight vector:

$$\boldsymbol{\beta}^0 = \mathbf{D}_0 {\mathbf{H}_0}^T \mathbf{T}_0, \tag{3}$$

where $\mathbf{D}_0 = ({\mathbf{H}_0}^T \mathbf{H}_0)^{-1}$ and $\mathbf{T}_0 = [t_1, t_2, \ldots, t_{N_0}]^T$.

(3) Set $k = 0$ 24.

Step 2 (sequential learning phase). Consider the following:

(1) Learn the $(k+1)$th training data: $d_{k+1} = (\mathbf{x}_{N_0+k+1}, t_{N_0+k+1})$.

(2) Calculate the partial hidden layer output matrix:

$$\mathbf{H}_{k+1} = \left[g\left(\mathbf{w}_1 \cdot \mathbf{x}_{N_0+k+1} + b_1\right) \ \cdots \ g\left(\mathbf{w}_L \cdot \mathbf{x}_{N_0+k+1} + b_L\right) \right]_{1\times L}. \tag{4}$$

Set $\mathbf{T}_{k+1} = [t_{N_0+k+1}]^T$.

(3) Calculate the output weight vector

$$\begin{aligned} \mathbf{D}_{k+1} &= \mathbf{D}_k - \mathbf{D}_k {\mathbf{H}_{k+1}}^T \left(\mathbf{I} + \mathbf{H}_{k+1} \mathbf{D}_k {\mathbf{H}_{k+1}}^T\right)^{-1} \mathbf{H}_{k+1} \mathbf{D}_k, \\ \boldsymbol{\beta}^{k+1} &= \boldsymbol{\beta}^k + \mathbf{D}_{k+1} {\mathbf{H}_{k+1}}^T \left(\mathbf{T}_{k+1} - \mathbf{H}_{k+1} \boldsymbol{\beta}^k\right). \end{aligned} \tag{5}$$

(4) Set $k = k + 1$. Go to Step 2(1).

4. Online Sequential Extreme Learning Machine Based on Granulation Division and SMOTE

To improve the classification accuracy of minority class, we proposed a new algorithm based on granulation division and SMOTE using extreme learning machine. The main idea is improving the accuracy of minority class and reducing the information loss of majority class.

For the convenience of description of the algorithm, we give some definitions in the beginning. Suppose that $D = \{(x_i, t_i), i = 1, 2, \ldots, N_1\}$ and $S = \{(y_i, t_i), i = 1, 2, \ldots, N_2\}$ represent majority sample set and minority sample set, respectively, where x_i and y_i mean m-dimensional vector. Dimension indicates the number of features. $t_i = 1$ means the corresponding sample is the majority and $t_i = 0$ means minority sample.

Definition 1 (maximum radius of granule). In the jth granule, the distance between the granule core and the furthest point is called maximum radius of granule of the jth granule:

$$R_j = \max_{x_{ij} \in X} \left(|C - X|\right), \tag{6}$$

where $C = (c_1, c_2, \ldots, c_{k_1})$, $j = 1, 2, \ldots, k_1$, and c_j is the coordinate of granule core.

Definition 2 (granule dispersion). Granule dispersion represents the discrete degree in a granule. Obviously, the granule dispersion is inversely proportional to the number of samples in a granule and is directly proportional to the maximum radius of granule:

$$\begin{aligned} 1 \text{ Separate}\left(c_j\right) &= \frac{R_j * \sum_{i=1}^{k_p} \left|c_j - x_{ij}\right|}{k_p^3} \\ &= \frac{R_j * \sum_{i=1}^{k_p} \sqrt{c_j^2 - 2x_{ij}c_j + x_{ij}^2}}{k_p^3}, \end{aligned} \tag{7}$$

where x_{ij} is the sample in the granule with c_j as the granule core. It is easy to know that the bigger the granule dispersion is, the more sparse and scattered the samples in the granule are and thus the higher the information loss is upon using granule core instead of the whole granule.

Definition 3 (sample weight). For each sample in the granule except the samples farthest from the granule core, the oversampling is conducted using SMOTE. Obviously, the sample weight is inversely proportional to the distance between the granule core and the virtual samples:

$$w_i = \left\| 1 - \frac{\sqrt{c_j^2 - 2x_{ij}c_j + x_{ij}^2}}{R_j} \right\|. \tag{8}$$

4.1. Offline Stage. Firstly, we refactor the imbalanced sample set using the proposed method and get the balanced sample set $A = \{(x_i, t_i) \mid i = 1, 2, \ldots, N_0\}$. Then establish the initial model. The main idea is undersampling for majority class by choosing all the granule cores, which can reduce the number of majority samples and ensure that the samples' distribution trend is consistent with the trend before undersampling.

For initial majority sample set $D = \{(x_i, t_i), i = 1, 2, \ldots, N_1\}$, the first granulation division is conducted. Then, we obtain the new majority sample set $D_n = \{(x_i, t_i), i = 1, 2, \ldots, N_{11}\}$.

Clustering algorithm [21] is adopted to simulate the process of granulation division. We set the clustering algorithm k_1 in the first granulation division according to the overall distribution of original samples. The up and down threshold values of maximum radius of granule are set to $[\eta_1, \eta_2]$, which can guarantee that the sample distribution trend keeps unchanged before and after the first undersampling. Then, we choose k_1 clustering center as the granule core and replace the original majority samples. Merging the new majority sample set $D_n = \{(x_i, t_i), i = 1, 2, \ldots, N_{11}\}$ and the minority sample set $S = \{(y_i, t_i), i = 1, 2, \ldots, N_2\}$, we obtain the new training sample set $A = \{(x_i, t_i) \mid i = 1, 2, \ldots, N_0\}$.

Given the hidden active function $g(x)$ and the number of hidden nodes L, choose input weight w_i and bias b_i, $i = 1, 2, \ldots, N_0$ randomly and calculate the hidden layer output matrix $\mathbf{H}_0$:

$$\begin{aligned}\mathbf{H}_0 &= \begin{bmatrix} h(x_1) \\ h(x_2) \\ \vdots \\ h(x_{N_0}) \end{bmatrix} \\ &= \begin{bmatrix} g(w_1 \cdot x_1 + b_1) & \cdots & g(w_L \cdot x_1 + b_L) \\ g(w_1 \cdot x_2 + b_1) & \cdots & g(w_L \cdot x_2 + b_L) \\ \vdots & & \\ g(w_1 \cdot x_{N_0} + b_1) & \cdots & g(w_L \cdot x_{N_0} + b_L) \end{bmatrix}.\end{aligned} \tag{9}$$

The output vector is $\mathbf{T}_0 = [t_1, t_2, \ldots, t_{N_0}]^T$ and the output weight is

$$\boldsymbol{\beta}_0 = {\mathbf{H}_0}^{\dagger}\mathbf{T}_0, \tag{10}$$

where ${\mathbf{H}_0}^{\dagger} = ({\mathbf{H}_0}^T\mathbf{H}_0)^{-1}{\mathbf{H}_0}^T$. Set $P_0 = ({\mathbf{H}_0}^T\mathbf{H}_0)^{-1}$, and then we have ${\mathbf{H}_0}^{\dagger} = P_0{\mathbf{H}_0}^T$.

Set $k = k + 1$.

4.2. Online Stage. Suppose that the sequential samples chunk is $\Omega_{k+1} = \{(x_i, t_i) \mid i = N_0 + 1 + k, \ldots, N_0 + 1 + k + \text{Block}\}$ in $(k + 1)$th step, where Block means the number of samples in Ω_{k+1}. Ω_{k+1} is divided into majority class Ω_d and minority class Ω_s according to the value of label t_i.

(1) Granulation division for majority classis Ω_d is conducted. We choose k_1 points as the initial granule core uniformly according to the samples distribution trend, where k_1 is set as about three times the number of minority samples. We can obtain k_1 clustering center in each iteration, namely, granule core, by equation of clustering algorithm $\mu_i^{(j+1)} = (1/n_i)\sum_{i=1}^{n_i} p_{it}$, until the distance between each sample and the clustering center meets the condition of the following equation. Finally, we can obtain k_1 clustering center $C = \{c_1, c_2, \ldots, c_{k_1}\}$, namely, k_1 granule core. Now, the majority sample set is Ω_{dn} and the imbalance rate reduces to 3 : 1:

$$\min \quad \sum \sum \operatorname{dist}(c_i, x)^2. \tag{11}$$

(2) Granulation computing for minority class: the value of k_2 is set to half the number of original minority samples:

$$k_2 = \frac{1}{2}\operatorname{size}(\Omega_s). \tag{12}$$

If there are other samples except the granule core, we add the virtual samples within the granule and granule boundary using SMOTE. Then, we obtain the new minority sample set by merging the virtual samples and original minority samples. The detailed description is as follows.

Step 1. Choose the granule core as the center of a circle. We add virtual samples between the granule core and all the other samples using SMOTE, as shown in the following equation:

$$\begin{aligned} x'_{jim} &= c_j + \operatorname{diff}_i \times \operatorname{rand}[0, 1] \\ &= c_j + (c_j - x_{ji}) \times \operatorname{rand}[0, 1]. \end{aligned} \tag{13}$$

The virtual samples are generated between the granule core and other samples in the granule. Every time there are m virtual samples generated. The value of m_1 is set according to the actual situation.

Step 2. According to the following equation:

$$\begin{aligned} x'_{ji(m+1)} &= c_j + \operatorname{diff}_i \times \operatorname{rand}[1, n_1] \\ &= c_j + (c_j - x_{ji}) \times \operatorname{rand}[1, n_1], \end{aligned} \tag{14}$$

m_2 virtual samples will be generated between the granule core and most of the samples in the granule, except the sample farthest from the granule core, which can ensure that the new virtual samples are not too far from the granule core and thus maintain credibility. Usually we set $m_2 \leq (1/3)m_1$, where m_1 means the total sample numbers in the granule. The new virtual samples expand the distribution range of the granule and do not affect the overall credibility simultaneously. Besides, the random number is between 1 and n_1, $1 < n_1 \leq 1.5$, which can ensure that the distance between new virtual samples and the granule core is farther than the distance between raw samples and the granule core, and most of the virtual samples are still in the granule.

Step 3. Set the sample weight of virtual samples according to Definition 3 and then update the virtual samples. Merging the virtual samples and the original minority samples, we can get the new minority sample set Ω_{sn}.

(3) The final balanced sample set is $\text{new}\Omega_{k+1} = \{(x_i, t_i) \mid i = N_0 + 1 + k, \ldots, N_0 + 1 + k + \text{newBlock}\}$, where newBlock is the number of samples in $(k + 1)$th step. Now the rate of majority class and minority class is probably between 1.5 : 1 and 1.1 : 1.

The corresponding hidden layer matrix of $\text{new}\Omega_{k+1}$ is $H_\Omega = [h_{k+N_0+1} \; h_{k+N_0+2} \; \cdots \; h_{k+N_0+\text{newBlock}}]$, and now the hidden layer matrix becomes $\mathbf{H}_{k+1} = [{H_k}^T {H_\Phi}^T]^T$. Update the network weight according to the following equation:

$$\boldsymbol{\beta}_{k+1} = {\mathbf{H}_{k+1}}^{\dagger}T_{k+1}, \tag{15}$$

where $T_{k+1} = [T_k{}^T T_\Omega{}^T]^T$ is the output vector and $\mathbf{H}_{k+1}{}^\dagger$ is $(\mathbf{H}_{k+1}{}^T\mathbf{H}_{k+1})^{-1}\mathbf{H}_{k+1}{}^T$. Let $P_{k+1} = (\mathbf{H}_{k+1}{}^T\mathbf{H}_{k+1})^{-1}$. We have

$$\mathbf{H}_{k+1}{}^\dagger = P_{k+1}\mathbf{H}_{k+1}{}^T, \tag{16}$$

because

$$\begin{aligned}\mathbf{H}_{k+1}{}^T\mathbf{H}_{k+1} &= \left[H_k{}^T H_\Phi{}^T\right]\left[H_k{}^T H_\Phi{}^T\right]^T \\ &= H_k{}^T H_k + H_\Phi{}^T H_\Phi;\end{aligned} \tag{17}$$

namely,

$$P_{k+1}{}^{-1} = P_k{}^{-1} + H_\Phi{}^T H_\Phi. \tag{18}$$

Calculate the inversion of both ends of the equation according to Sherman-Morrison matrix inversion lemma. We obtain the recursive expression of P_{k+1}:

$$P_{k+1} = \left(P_k{}^{-1} + H_\Phi{}^T H_\Phi\right)^{-1} = P_k - \frac{P_k H_\Phi{}^T H_\Phi P_k}{I + H_\Phi P_k H_\Phi{}^T}. \tag{19}$$

So P_{k+1} can be calculated based on P_k, which reduces calculation and greatly improves the computational efficiency. We can obtain $\mathbf{H}_{k+1}{}^\dagger$ by substituting (15) into (14) and then update the network weight $\boldsymbol{\beta}_{k+1}$.

5. The Reliability Analysis

According to the discussion as above, we reduce the majority samples using granulation division both in offline stage and in online stage. For the original majority sample set $\theta_d = \{(x_i, t_i),\ i = 1, 2, \ldots, M\}$, we only choose m most representative samples and get the new balanced majority sample set θ'_d. Although the imbalanced phenomenon could be reduced to some extent, there is a loss of information in the undersampling because of abandoning some samples. To illustrate the rationality of the proposed method, we give the lower bound of the model reliability after undersampling based on information entropy [22], which can indicate indirectly that there is upper bound of the information loss in undersampling.

Suppose that the loss sample set is $A = \{(x_{ij}, t_i),\ j = 1, 2, \ldots, m\}$ in every online undersampling, where x_{ij} is a sample in the granule centered with the granule core c_j. As discussed in Section 4.2, we reject all samples in this granule except c_j. The sample weight of c_j is defined as follows:

$$\omega\left(c_j\right) = \frac{1}{\text{Separate}\left(c_j\right)} = \frac{k_p^3}{R_j * \sum_{i=1}^{k_p}\left|c_j - x_{ij}\right|}. \tag{20}$$

Then the missed classed probability is

$$P_f = \sum_{j=1}^{m} w_k = \sum_{j=1}^{m}\left(1 - \frac{k_p^3}{R_j * \sum_{i=1}^{k_p}\left|c_j - x_{ij}\right|}\right), \tag{21}$$

where k_p means the number of samples and $\sum_{i=1}^{k_p}|c_j - x_{ij}|$ means the sum of Euclidean distance between each sample and the granule core.

In offline stage, the loss sample set is $A_1 = \{(y_{ij1}, t_i),\ j = 1, 2, \ldots, m_0\}$, where y_{ij1} means the sample in the jth granule. The granule core c_{j1} will join the final balanced sample set representing the whole granule. So the misclassification rate is

$$P_{f1} = \sum_{l=1}^{m_0} w_{k1} = \sum_{l=1}^{m_0}\left(1 - \frac{k_{p1}^3}{R_{j1} * \sum_{i=1}^{k_{p1}}\left|c_{j1} - x_{ij1}\right|}\right). \tag{22}$$

Theorem 4. *At present, N is the sample number of majority set D and L is the number of misclassification majority samples. Let R_L represent the lower bound of model reliability. Because binary classification result obeys the binomial distribution, the lower bound of model reliability can be obtained when the confidence coefficient is determined:*

$$\begin{aligned}\sum_{r=0}^{L}\left(\frac{N}{r}\right) R_L^{N-r}\left(1 - R_L\right)^r &= \sum_{r=0}^{N*P_f}\left(\frac{N}{r}\right) R_L^{N-r}\left(1 - R_L\right)^r \\ &= \sum_{r=0}^{N*\sum_{j=1}^{m}(1-k_p^3/(R_j*\sum_{i=1}^{k_p}|c_j-x_{ij}|))}\left(\frac{N}{r}\right) R_L^{N-r}\left(1 - R_L\right)^r \\ &= 1 - \alpha,\end{aligned} \tag{23}$$

where L is negatively correlated with R_L. It can be seen that the fuzzy reliability is only related to the discrete degree.

Proof. According to the definition of fuzzy reliability $\sum_{r=0}^{F}(N/r)R_L^{N-r}(1 - R_L)^r = 1 - \alpha$, R_L reaches the maximum when F is the minimum with definite α, because

$$\begin{aligned}F &\le L = N * P_f \\ &= N * \sum_{j=1}^{m}\left(1 - \frac{k_p^3}{R_j * \sum_{i=1}^{k_p}\left|c_j - x_{ij}\right|}\right).\end{aligned} \tag{24}$$

As can be seen from the above equation, R_L is only related to $k_p^3/(R_j * \sum_{i=1}^{k_p}|c_j - x_{ij}|)$. We know $k_p^3/(R_j * \sum_{i=1}^{k_p}|c_j - x_{ij}|) = 1/\text{Separate}(c_j)$. So the smaller the maximum radius of granule is, the smaller the distance sum of samples is, the higher the number of samples in granule is, the smaller the dispersion is, and the bigger the value of $k_p^3/(R_j * \sum_{i=1}^{k_p}|c_j - x_{ij}|)$ is, which will cause smaller lower bound of reliability and more reliable model. □

Theorem 5. *The sample size of majority set D_0 is N_0 in offline stage and the number of misclassification samples is L_0. The reliability can be obtained according to the following equation. And the reliability is only related to the value of $k_{p1}^3/(R_{j1} * \sum_{i=1}^{k_{p1}}|c_{j1} - x_{ij1}|)$:*

$$\begin{aligned}\sum_{r=0}^{L_0}\left(\frac{N_0}{r}\right) R_{L_0}^{N_0-r}\left(1 - R_{L_0}\right)^r &= \sum_{r=0}^{N_0*P_{f1}}\left(\frac{N_0}{r}\right) \\ &\cdot R_{L_0}^{N_0-r}\left(1 - R_{L_0}\right)^r\end{aligned}$$

$$= \sum_{r=0}^{N_0 * \sum_{l=1}^{m_0} (1-k_{p1}^3/(R_{j1} * \sum_{i=1}^{k_{p1}} |c_{j1}-x_{ij1}|))} \left(\frac{N}{r}\right) \cdot R_{L_0}^{N_0-r} \left(1 - R_{L0}\right)^r = 1 - \alpha. \tag{25}$$

Proof. According to the definition of fuzzy reliability, we have

$$\begin{aligned} F &\leq L_0 = N_0 * P_{f1} \\ &= N_0 * \sum_{l=1}^{m_0} \left(1 - \frac{k_{p1}^3}{R_{j1} * \sum_{i=1}^{k_{p1}} \left|c_{j1} - x_{ij1}\right|}\right). \end{aligned} \tag{26}$$

According to the equation, the value of R_{L_0} is related to the sum of distance between the granule core and other samples. The smaller the distance sum is, the smaller the dispersion is and the more compact the samples in granule are, which will cause the bigger fuzzy reliability and more reliable model. □

Theorems 4 and 5 prove the reasonability of the proposed algorithm from the point of information entropy. Considering the extreme case, if the granule dispersion is 0, namely, not undersampling by granulation division, the misclassification rate of majority class is almost 0; that is, $\lim_{P_s \to 1} H(\Phi) \to 0$ means that it does not provide the information entropy and the information loss is 0, which is accordant with the practical situation.

6. Simulation Experiment

In order to demonstrate the effectiveness and the superiority of the proposed algorithm, we conduct the simulation experiments on the chessboard-shaped dataset with uneven density distribution and the imbalanced distribution meteorological data of Macao in 2010 and 2012. At the same time, we compare the experimental result of our algorithm with that of SVM (support vector machine) [23], OS-ELM (online sequential ELM) [14], and MCOS-ELM (Metacognitive OS-ELM) [16]. Among them, MCOS-ELM is an online extreme learning algorithm presented by Vong et al. [16] for the online data imbalance problem. For better demonstration, we call the proposed method DGSMOTE (Division of Granulation and SMOTE OS-ELM). Before the training, we apply the normalization procedure to the dataset. We take the average value of 30 trials as the final experimental result.

6.1. Construct the Chessboard-Shaped Dataset. In the chessboard-shaped dataset, both the majority and the minority samples take up eight cells in the chessboard. According to the respective class in each cell, some data chunks are randomly generated. Ultimately, the quantity of majority and minority samples is 1000 ∗ 8 and 100 ∗ 8, respectively; that is to say, the ratio of the classes is 10 : 1. The testing data are generated with the same method.

6.1.1. Experimental Results Analysis of Chessboard-Shaped Data. In the offline stage, we conduct the undersampling on the majority samples first. Then the changes in the distribution of the chessboard-shaped data are shown as Figure 1(b) after the first granulation division. In the online stage, after conducting granulation division, we process the SMOTE algorithm to realize the oversampling for the minority samples. As a result, the changes of the dataset are shown as Figure 1(c).

Table 1: Changes in the numbers before and after balancing the offline data.

Dataset	Before		After	
	Majority	Minority	Majority	Minority
Chessboard-shaped data	7997	807	2500	2449

Table 2: Comparative results on chessboard-shaped dataset.

Algorithms	DGSMOTE	OS-ELM	MCOS-ELM	LS-SVM
Testing time (s)	0.0874	0.1976	0.0531	132.476
Minority training accuracy	0.9167	0.6329	0.7438	0.6385
Majority training accuracy	0.9414	0.8988	0.8051	0.9792
Whole testing accuracy	0.9363	0.8439	0.8612	0.9606
G-mean	0.9325	0.7334	0.7445	0.7822

It can be seen from Figure 1 that, compared with the original samples, now the classifications of dataset at the moment are nearly balanced. For the different classes of the chessboard-shaped dataset, Table 1 presents the changes in their numbers before and after the process of this proposed algorithm.

In the simulation experiment, the activation function of the hidden layer is set as "sig" and the numbers of hidden nodes are set as 140. We take the mean value of 30 trials as the experimental result. Finally, the performance comparison of the four models is shown as Table 2.

From Table 2, though the DGSMOTE's whole testing accuracy is not the highest among the models, its testing accuracy and testing times do not appear to be much different from the other three algorithms. In addition, the minority training accuracy of DGSMOTE is much higher than that of others. Compared with LS-SVM, DGSMOTE shows superior performance in both testing training speed and accuracy. This demonstrates the instantaneity of the proposed algorithm. Furthermore, the new DGSMOTE presents good performance upon using *G*-mean to evaluate specialty and the sensitivity of the algorithm. It can highly improve the classification accuracy of minority with less decrease of amplitude of majority accuracy. At the same time, it can eliminate the bias which is generated by applying traditional algorithm to handle the imbalanced data.

To strengthen the reliability and observability of our algorithm, the classification results and the minority accuracy variation with different numbers of hidden nodes on chessboard-shaped dataset are shown in Figures 2 and 3, where the dark spot means the misclassified samples. They

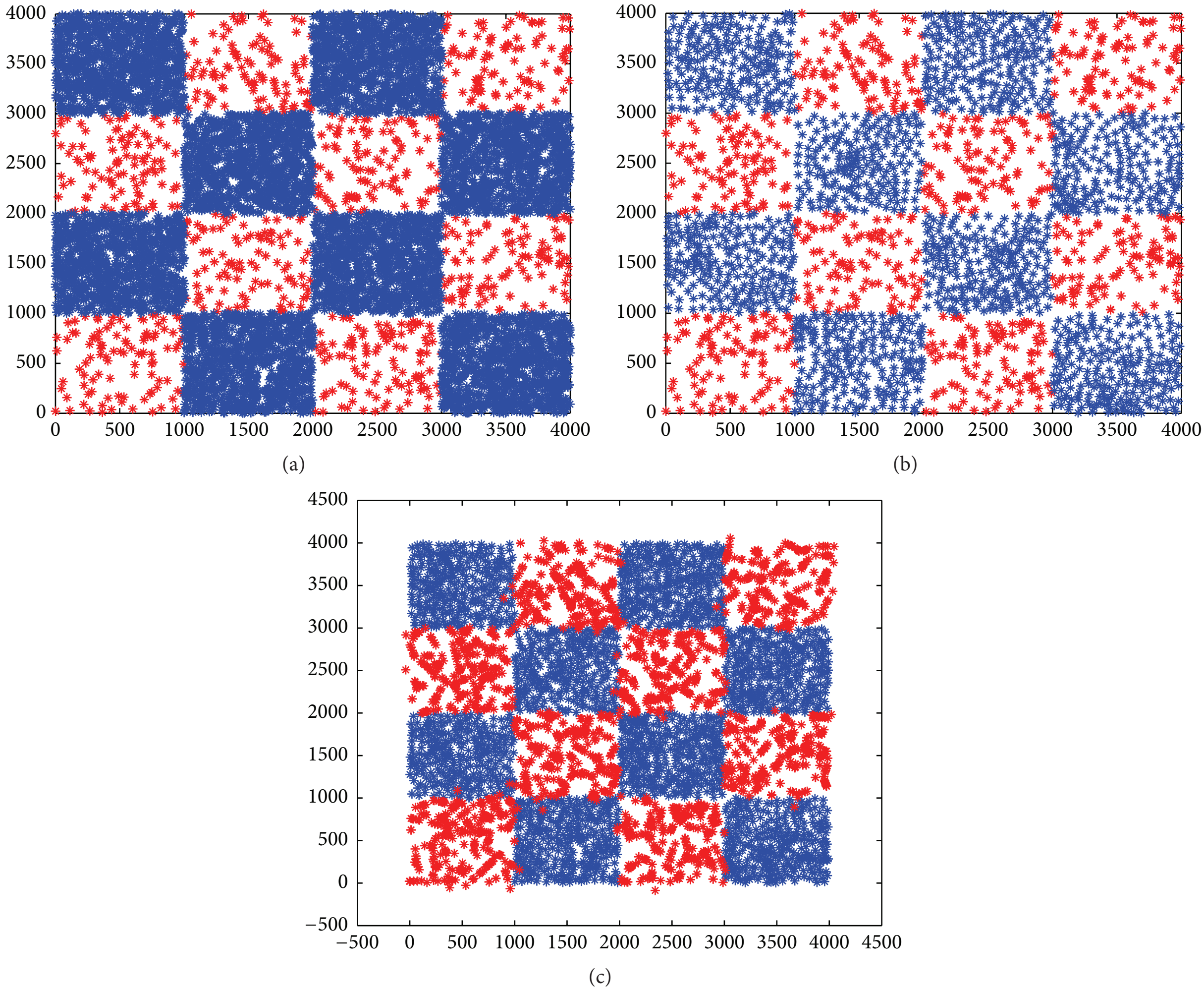


Figure 1: The distribution of the offline chessboard-shaped data (a) before and (b) after the first granulation and (c) the result after using SMOTE.

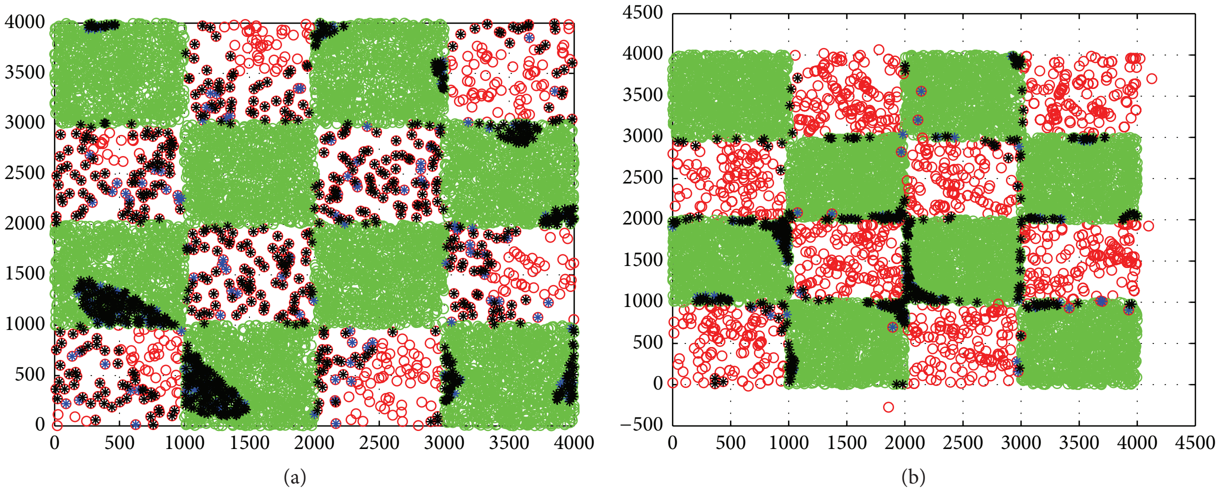


Figure 2: Classification results of on chessboard-shaped dataset. (a) OS-ELM and (b) DGSMOTE.

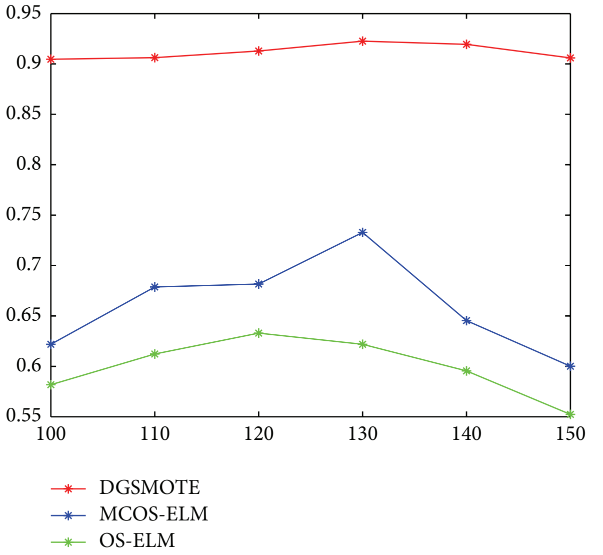


Figure 3: Classification accuracy on minority class with different number of hidden nodes for chessboard-shaped dataset.

also reflect the good generalization and learning performance of DGSMOTE.

From Figure 2, classification accuracy of minority samples is significantly less than that of majority samples upon using OS-ELM to classify the dataset. Namely, the model of OS-ELM possesses obvious bias in the classification. Compared with the improved OS-ELM algorithms, the DGSMOTE has a better overall classification performance with little effect on the majority classification accuracy. During the procedure of undersampling and oversampling in our algorithm, both the whole distribution characteristics and the original feature of the samples are fully considered. So loss information value of the balanced samples is low and stable. We transform the number of the hidden nodes for the three online extreme learning algorithms. Then the accuracy of each node for the corresponding algorithm is obtained by taking the mean value of 20 trials. Hence, we get Figure 3. By observing the changes of minority accuracy, we know that the whole performance of DGSMOTE algorithm is not only high but also stable.

After synthesizing all the above indicators, it is obvious that the DGSMOTE possesses effective generalization performance and outstanding learning ability. In order to display the sensitivity and specificity of our algorithm, we employ ROC curve to reveal the excellent performance. The ROC curves of the four models on the chessboard-shaped dataset are shown in Figure 4.

AUC denotes the area under the ROC curve. The larger the value of AUC is, the better the classification is. From Figures 1–3, we can know that the proposed DGSMOTE algorithm significantly outperforms the other three models with better overall performance and lower minority misclassification rate. Besides, it reduces the loss cost generated by misclassification because of its strong recognition capability and lower classification bias.

Table 3: Changes in the numbers before and after balancing the offline data.

Dataset	Before		After	
	Majority	Minority	Majority	Minority
Forecasting data in 2010	334	31	110	97
Forecasting data in 2011	330	35	110	104

Table 4: Comparative results on Macao forecasting data in 2010.

Algorithms	DGSMOTE	OS-ELM	MCOS-ELM	LS-SVM
Testing time (s)	0.0399	0.0468	0.0468	1.6224
Minority training accuracy	0.9655	0.5775	0.7037	0.7009
Majority training accuracy	0.9481	0.9771	0.9746	0.9286
Whole testing accuracy	0.9546	0.9432	0.9368	0.9025
G-mean	0.9472	0.7403	0.8213	0.7655

6.2. Experimental Results Analysis of Macao Forecasting Data. Macao forecasting dataset is obtained from the website of the Macao Meteorological Bureau [24]. Compared with chessboard-shaped dataset, it has less samples but more attributions and is a kind of flow distribution data. According to its own features, we choose PM_{10} and SO_2 as the two main characteristics from all the six features for illustration. And the changes after the first granulation division are shown in Figure 5(b).

After the first granulation division, the imbalanced ratio of the new sample set is markedly decreased compared to that of the original dataset.

In the online stage, we first conduct granular computing. Next, we apply SMOTE algorithm to process the oversampling for the minority. Figure 5(c) displays how the minority samples of Macao forecasting data change after the procedure of DGSMOTE.

It is obvious form Figure 5 that the sample data are nearly balanced after the oversampling and undersampling. Table 3 shows the changes in the number of the observations before and after being handled by our algorithm.

The next step is to use the new balanced sample set to establish the initial model of the online extreme learning machine. Similarly, the activation function of the hidden layer is set as “sig.” According to features of the forecasting data, the numbers of hidden nodes were assigned as 30. The four models established by the four algorithms conduct the learning on the two different forecasting datasets, respectively. Finally, the comparative performances of the models are presented in Tables 4 and 5.

As can be seen from Tables 4 and 5, compared with the other three algorithms, DGSMOTE can effectively increase

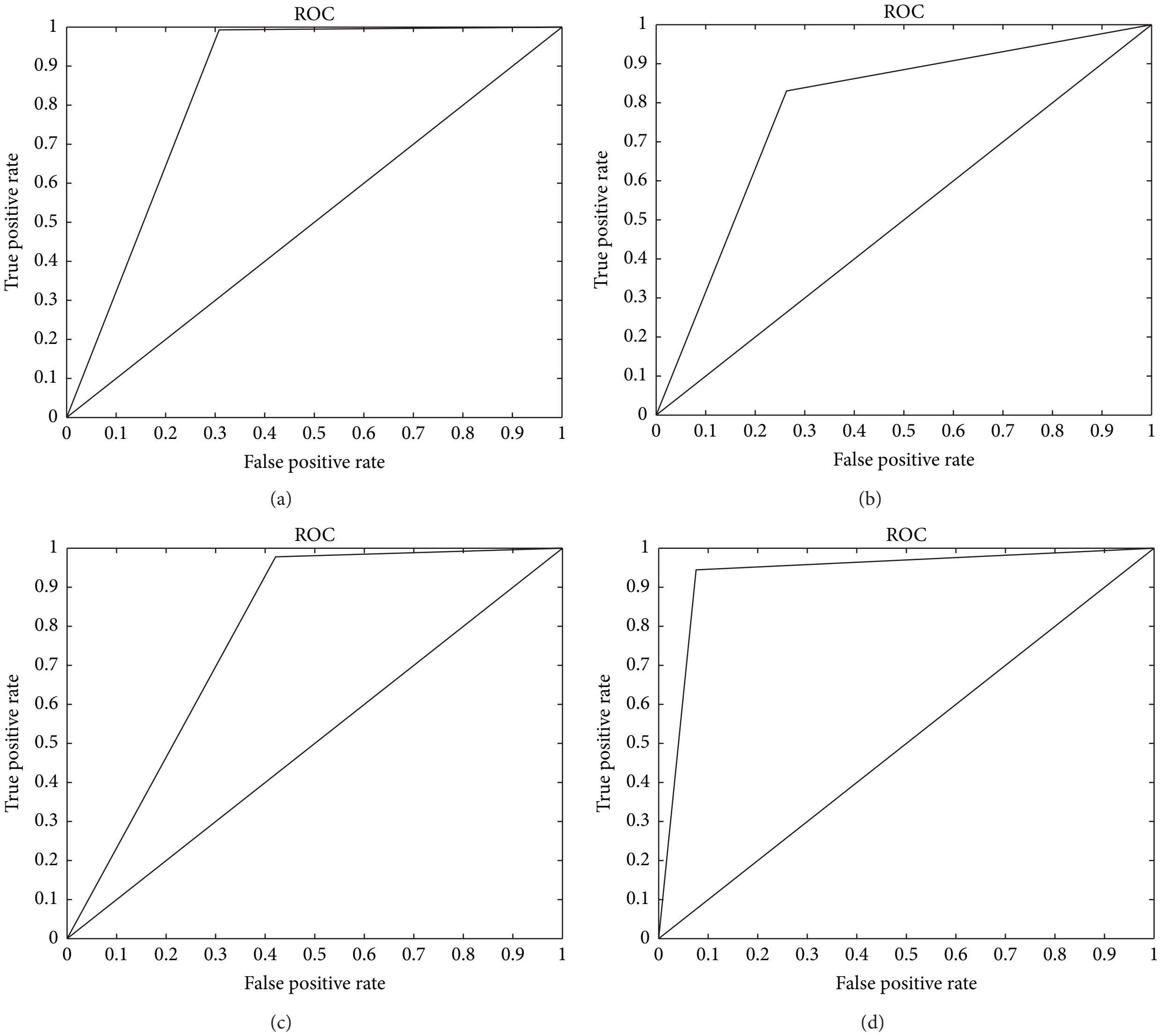


Figure 4: Comparison of the ROC on chessboard-shaped dataset for the four algorithms (a) SVM, (b) OS-ELM, (c) MCOS-ELM, and (d) DGSMOTE.

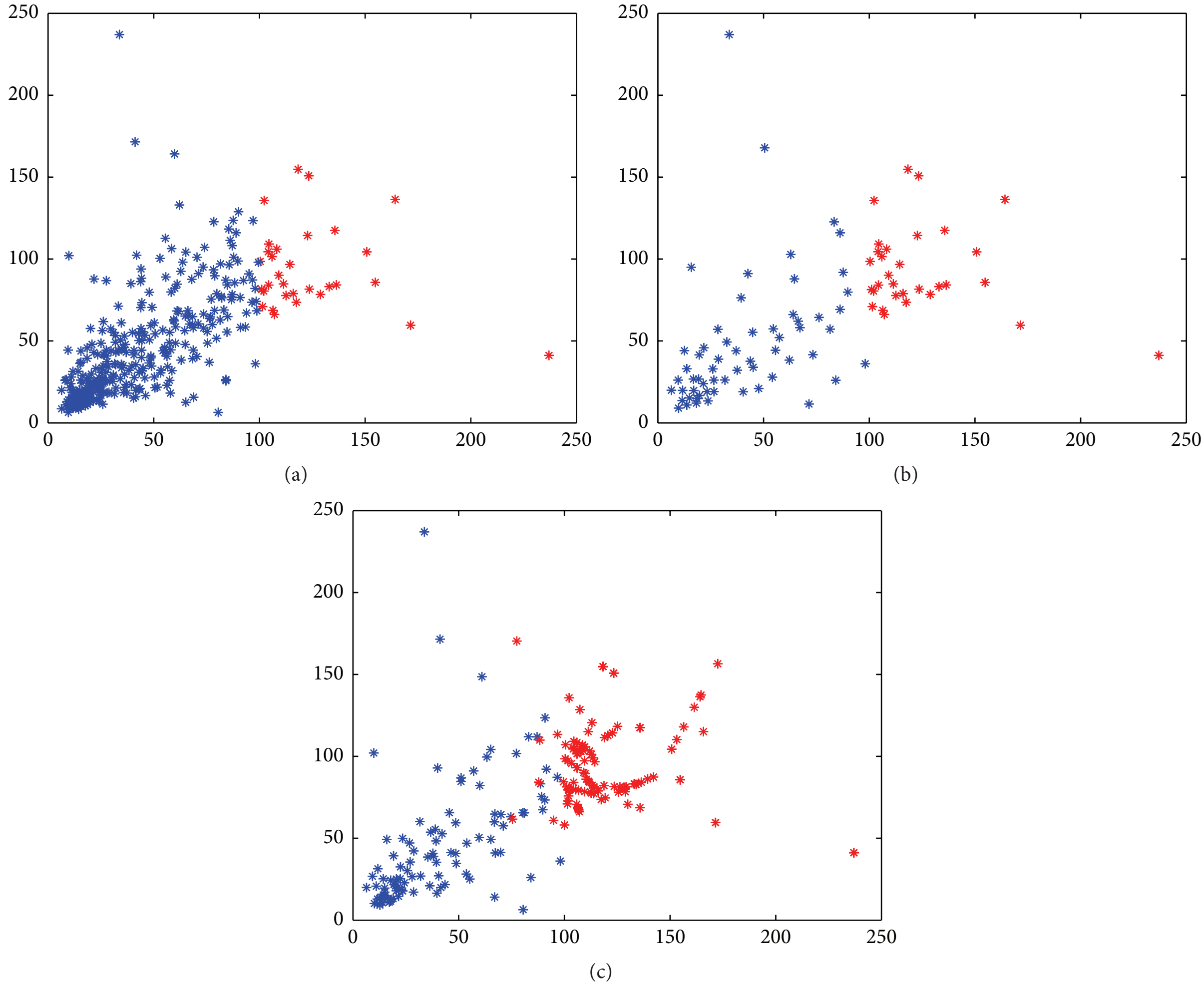
Table 5: Comparative results on Macao forecasting data in 2011.

Algorithms	DGSMOTE	OS-ELM	MCOS-ELM	LS-SVM
Testing time (s)	0.0321	0.0412	0.0425	1.638
Minority training accuracy	0.9469	0.5723	0.7206	0.7457
Majority training accuracy	0.9456	0.9756	0.9673	0.9492
Whole testing accuracy	0.9495	0.9421	0.9662	0.9263
G-mean	0.9347	0.7387	0.8399	0.7948

the classification accuracy and recognition capability of the minority samples upon dealing with the practical imbalanced problem. However, the overall testing accuracy and the value of G-mean suffer no decline. This case declares that most original features of the minority samples are reserved after the undersampling. It also reflects the low information loss and the validity of the granulation division. Meanwhile, the shorter testing time reveals the fast learning speed, strong recognition ability, and good instantaneity of the DGSMOTE algorithm.

In order to make the validity and stability of our proposed algorithm more clear, the classification results and the minority accuracy variation with different numbers of hidden nodes on the two datasets are shown in Figures 6, 7, and 8, respectively. In Figures 6 and 7, the dark spots mean the misclassified samples.

From Figures 6 and 7, our proposed algorithm has superior recognition ability to the other three algorithms as well as avoiding significant decline of majority class accuracy.

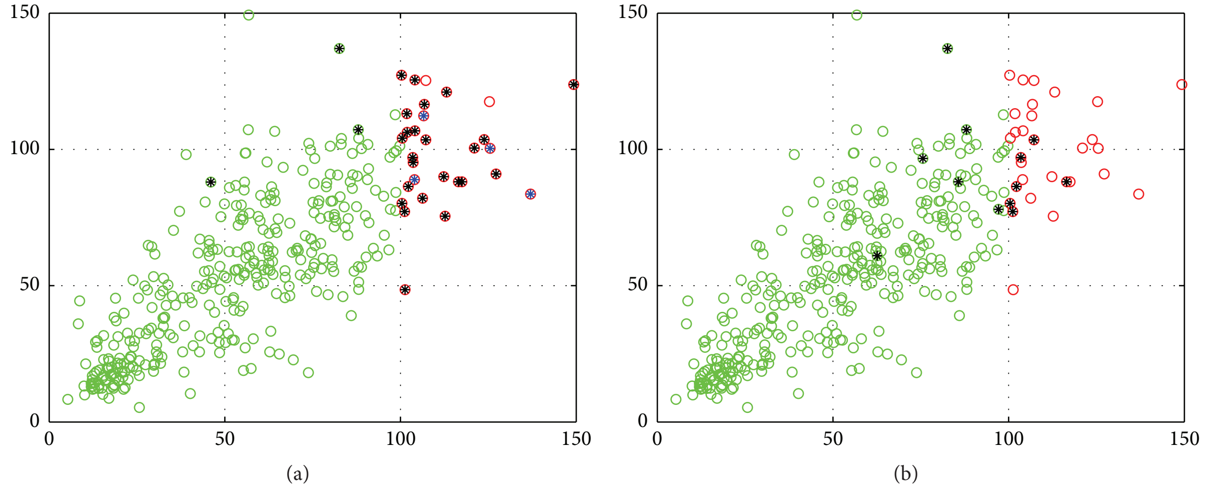


(a)

(b)

(c)

Figure 5: The distribution of the forecasting data (a) before and (b) after the first granulation division and (c) the result after using SMOTE.



(a)

(b)

Figure 6: Classification results of (a) OS-ELM and (b) DGSMOTE on forecasting data in 2010 year.

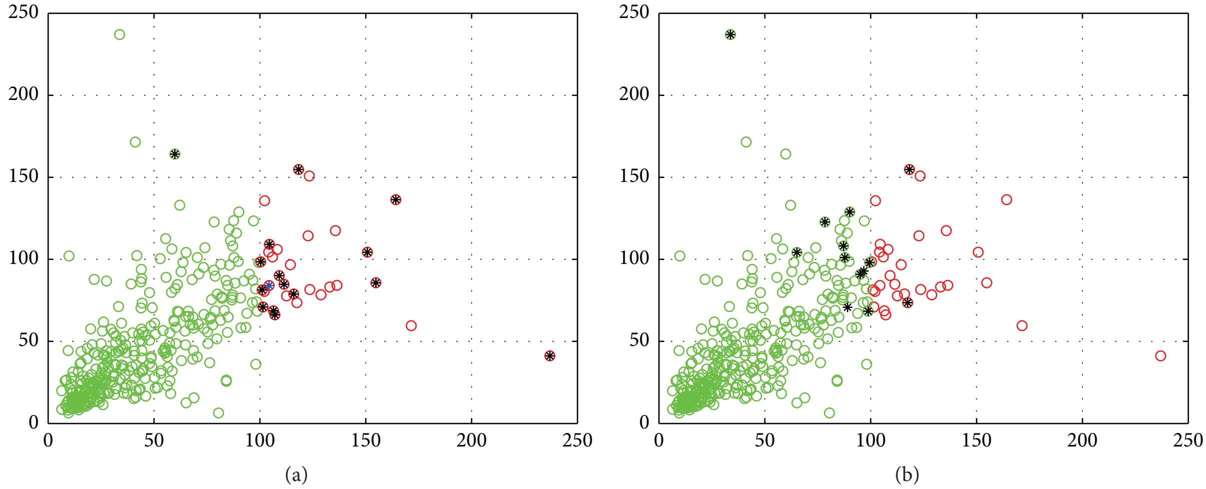


Figure 7: Classification results of (a) OS-ELM and (b) DGSMOTE on forecasting data in 2011 year.

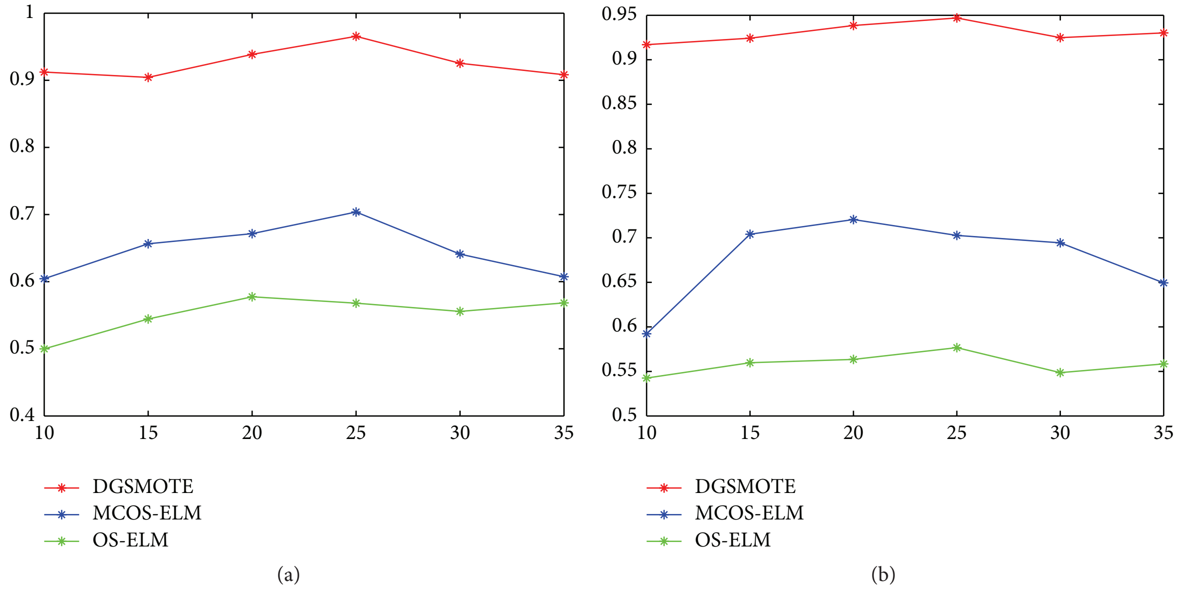


Figure 8: Classification accuracy on minority class with different number of hidden nodes on Macao forecasting data in (a) 2010 year and (b) 2011 year.

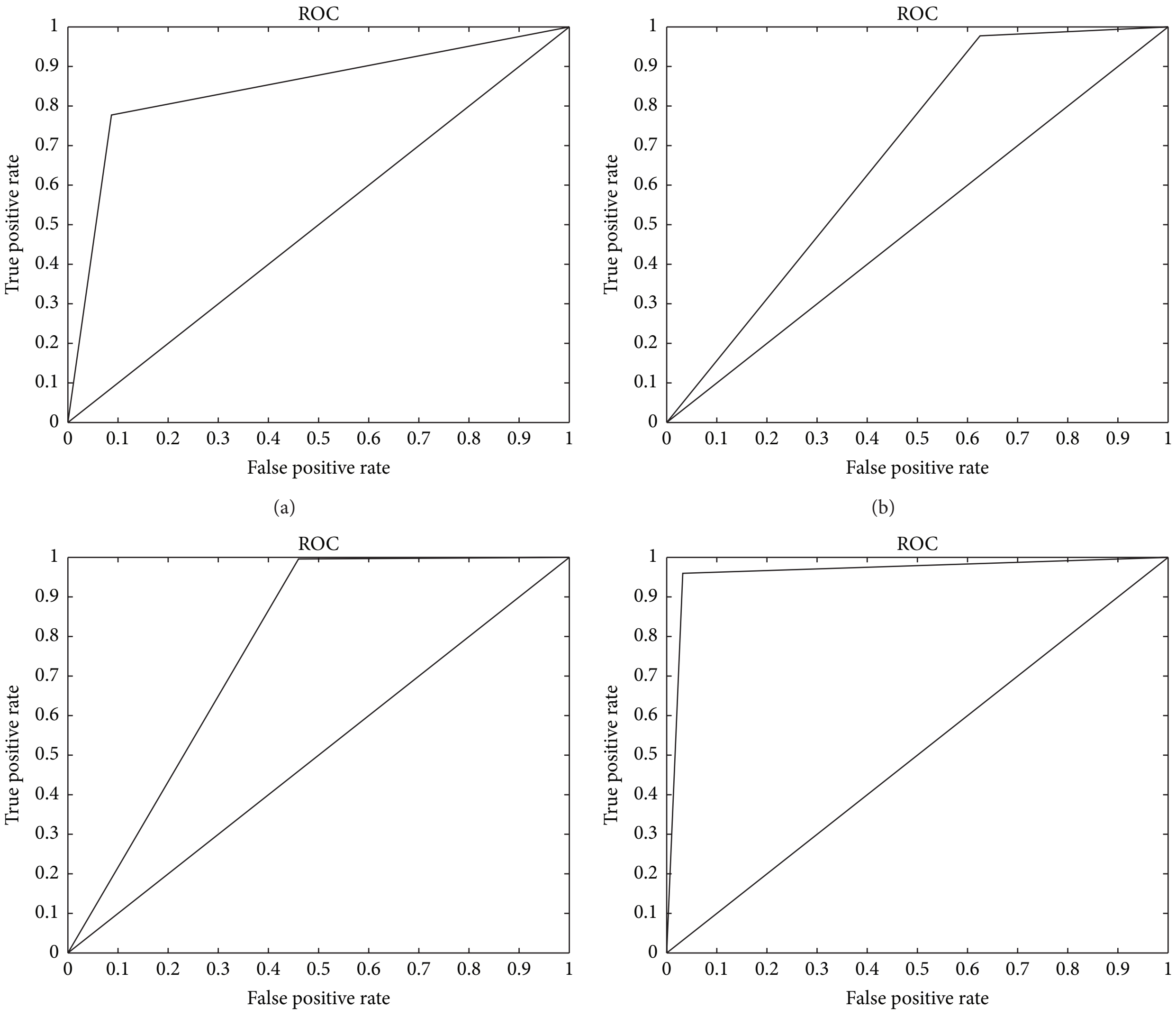
That is to say, the DGSMOTE effectively eliminates the bias generated by applying the original OS-ELM algorithms to handle the imbalanced problems. In Figure 8, the curves of DGSMOTE are much smoother without erratic fluctuation along with the variation of the number of hidden nodes. And they further testify the favorable generalization and stability of our proposed algorithm.

We still use ROC curves to exhibit the outstanding overall effect and superior performance of the proposed algorithm. Figures 9 and 10 indicate the ROC curves of the four models on the Macao forecasting data in 2010 year and 2011 year, respectively.

According to ROC curves in Figures 9 and 10, it is obvious that the DGSMOTE algorithm has an advantage over the other three algorithms and possesses better overall performance and recognition ability upon dealing with the flow distribution imbalanced data. This shows more research and application value for the practical problems.

7. Conclusion

In this paper, a novel classification approach based on SMOTE is proposed from the application in actual engineering. In the offline stage, we conduct the granulation



(a)

(b)

(c)

(d)

Figure 9: Comparison of the ROC on Macao forecasting data in 2010 year for the four algorithms (a) SVM, (b) OS-ELM, (c) MCOS-ELM, and (d) DGSMOTE.

division according to the distribution and the clustering characteristics of the majority samples. The central sample in each granule is used to replace the granule itself. Finally, the balanced offline dataset is obtained. In the online stage, we first process the granulation division for the minority class on the basis of the offline stage and then conduct the SMOTE to realize the oversampling of the minority samples. Our algorithm effectively increases the classification accuracy of the minority class under the premise that the overall distribution was unchanged and the information loss of majority samples reduced.

Furthermore, entropy theorem is used to testify the rationality of the proposed algorithm. The final experimental results demonstrate that the overall generalization performance, classification efficiency, and classification accuracy of the online imbalanced samples can get improved by applying granulation division to make the dataset balanced. For the online imbalanced small sample set and the large scale data, our research is of both great theoretical significance and practical value.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The authors would like to acknowledge the author W. Mao of [18] for useful discussion and Ph.D. students B. Chen and H. Fang for their work in simulation experiment. This work was supported by Fundamental Research Funding of

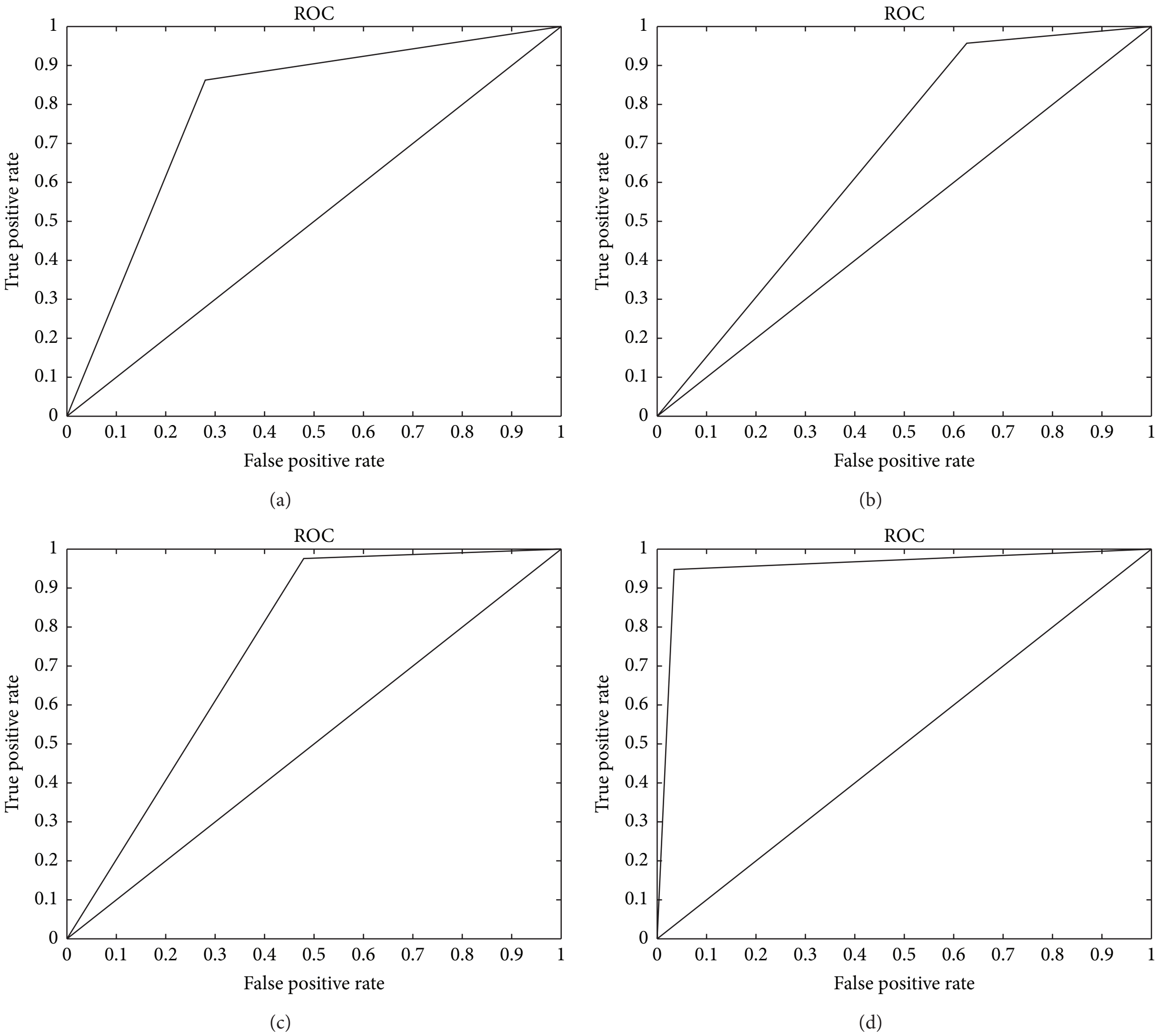


Figure 10: Comparison of the ROC on Macao forecasting data in 2011 year for the four algorithms (a) SVM, (b) OS-ELM, (c) MCOS-ELM, and (d) DGSMOTE.

Northwestern Polytechnical University (no. JC20120215) and Defense Fundamental Research Funding of China.

References

[1] Z. Yang, L. Qiao, and X. Peng, “Research on data mining method for imbalanced dataset based on improved SMOTE,” *Acta Electronica Sinica*, vol. 12, no. 35, pp. 22–26, 2007.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.

[3] N. Verbiest, E. Ramentol, C. Cornelis, and F. Herrera, “Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection,” *Applied Soft Computing*, vol. 22, pp. 511–517, 2014.

[4] M. Gao, X. Hong, S. Chen, and C. J. Harris, “A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems,” *Neurocomputing*, vol. 74, no. 17, pp. 3456–3466, 2011.

[5] Z.-Q. Zeng, Q. Wu, B.-S. Liao, and J. Gao, “A classfication method for imbalance data set based on kernel smote,” *Acta Electronica Sinica*, vol. 37, no. 11, pp. 2489–2495, 2009.

[6] P. Jeatrakul, K. W. Wong, and C. C. Fung, “Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm,” in *Neural Information Processing*, vol. 6444, pp. 152–159, Springer, Berlin, Germany, 2010.

[7] W. Wang, Z. Xu, W. Lu, and X. Zhang, “Determination of the spread parameter in the Gaussian kernel for classification and regression,” *Neurocomputing*, vol. 55, no. 3-4, pp. 643–663, 2003.

[8] J. P. Hwang, S. Park, and E. Kim, “A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function,” *Expert Systems with Applications*, vol. 38, no. 7, pp. 8580–8585, 2011.

[9] H. Yu, C. Mu, C. Sun, W. Yang, X. Yang, and X. Zuo, “Support vector machine-based optimized decision threshold adjustment strategy for classifying imbalanced data,” *Knowledge-Based Systems*, vol. 76, pp. 67–78, 2015.

[10] P. Turney, "Types of cost in inductive concept learning," *The Computer Research Repository*, pp. 15–21, 2002.

[11] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Proceedings of the 15th European Conference on Machine Learning (ECML '04)*, vol. 3201 of *Lecture Notes in Computer Science*, pp. 39–50, September 2004.

[12] C. X. Zhang and J. S. Zhang, "A survey of selective ensemble learning algorithms," *Chinese Journal of Computers*, vol. 34, no. 8, pp. 1399–1410, 2011.

[13] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 2, pp. 513–529, 2012.

[14] N.-Y. Liang, G.-B. Huang, P. Saratchandran, and N. Sundararajan, "A fast and accurate online sequential learning algorithm for feedforward networks," *IEEE Transactions on Neural Networks*, vol. 17, no. 6, pp. 1411–1423, 2006.

[15] W. Zong, G.-B. Huang, and Y. Chen, "Weighted extreme learning machine for imbalance learning," *Neurocomputing*, vol. 101, no. 3, pp. 229–242, 2013.

[16] C.-M. Vong, W.-F. Ip, P.-K. Wong, and C.-C. Chiu, "Predicting minority class for suspended particulate matters level by extreme learning machine," *Neurocomputing*, vol. 128, pp. 136–144, 2014.

[17] W. Mao, J. Wang, and Z. Xue, "An ELM-based model with sparse-weighting strategy for sequential data imbalance problem," *International Journal of Machine Learning and Cybernetics*, 2016.

[18] W. Mao, Y. Tian, J. Wang et al., "Research on granular extreme learning machine for sequential imbalanced data," in *Proceedings of the 26th Chinese Process Control Conference (CPCC '15)*, Nangchang, China, 2015.

[19] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1–3, pp. 489–501, 2006.

[20] W. Mao, S. Zhao, X. Mu, and H. Wang, "Multi-dimensional extreme learning machine," *Neurocomputing A*, vol. 149, pp. 160–170, 2015.

[21] S. Zahra, M. A. Ghazanfar, A. Khalid, M. A. Azam, U. Naeem, and A. Prugel-Bennett, "Novel centroid selection approaches for KMeans-clustering based recommender systems," *Information Sciences*, vol. 320, pp. 156–189, 2015.

[22] L. Tang, B. Song, Z. Li, and K. Zheng, "A fuzzy reliability evaluation method for sub-sample products based on information entropy theory," *Journal of Projectiles, Rockets, Missiles and Guidance*, vol. 25, no. 1, pp. 214–216, 2005.

[23] W. Mao, J. Xu, C. Wang, and L. Dong, "A fast and robust model selection algorithm for multi-input multi-output support vector machine," *Neurocomputing*, vol. 130, pp. 10–19, 2014.

[24] SMG E-publication Download Page, 2013, http://www.smg.gov.mo/www/ccaa/pdf/e_pdf_download.php.

Advances in Operations Research
Advances in Decision Sciences
Journal of Applied Mathematics
Algebra
Journal of Probability and Statistics
The Scientific World Journal
International Journal of Differential Equations
International Journal of Combinatorics
Advances in Mathematical Physics
Journal of Complex Analysis
Journal of Mathematics
Mathematical Problems in Engineering
Abstract and Applied Analysis
Discrete Dynamics in Nature and Society
International Journal of Mathematics and Mathematical Sciences
Journal of Discrete Mathematics
Journal of Function Spaces
International Journal of Stochastic Analysis
Journal of Optimization