Computing and Informatics, Vol. 34, 2015, 1017–1037

OVERSAMPLING METHOD FOR IMBALANCED CLASSIFICATION

Zhuoyuan Zheng

Guangxi Key Laboratory of Trusted Software
Guilin University of Electronic Technology, Guilin 541004, China
e-mail: zhengzhuoyuan@163.com

Yunpeng Cai, Ye Li

Shenzhen Institutes of Advanced Technology
Key Laboratory for Biomedical Informatics and Health Engineering
Chinese Academy of Sciences
1068 Xueyuan Avenue, Shenzhen University Town, Shenzhen 518055
e-mail: {yp.cai, ye.li}@siat.ac.cn

Abstract. Classification problem for imbalanced datasets is pervasive in a lot of data mining domains. Imbalanced classification has been a hot topic in the academic community. From data level to algorithm level, a lot of solutions have been proposed to tackle the problems resulted from imbalanced datasets. SMOTE is the most popular data-level method and a lot of derivations based on it are developed to alleviate the problem of class imbalance. Our investigation indicates that there are severe flaws in SMOTE. We propose a new oversampling method SNOCC that can compensate the defects of SMOTE. In SNOCC, we increase the number of seed samples and that renders the new samples not confine in the line segment between two seed samples in SMOTE. We employ a novel algorithm to find the nearest neighbors of samples, which is different to the previous ones. These two improvements make the new samples created by SNOCC naturally reproduce the distribution of original seed samples. Our experiment results show that SNOCC outperform SMOTE and CBSO (a SMOTE-based method).

Keywords: Classification, imbalanced dataset, oversampling, SMOTE, SNOCC

Mathematics Subject Classification 2010: 68T10

1 INTRODUCTION

Classification is by far the most important machine learning topic. Various classification algorithms, such as decision tree, BP neural networks, Bayesian networks, k-nearest neighbor, support vector machine, etc., were developed and widely used in many fields. Almost all algorithms suffer from the problem of imbalanced dataset, in which there are more instances belonging to some classes than others. Imbalanced data usually causes biases in classification and leads to poor generalization performance. For many real-world applications, including fraud detection [1, 2], bioinformatics [3, 4, 5, 6], text classification [7, 8], medical field [9, 10], etc., the class of interest has a frequency of less than 0.01 among all cases. In these applications, the minority is most interesting and its identification is of utmost importance. This requires a fairly high detection rate of the minority class and usually allows a small error rate in the majority class since the cost of misclassifying a majority instance can be relatively low. The class imbalance problem is of the crucial importance. It can cause a significant bottleneck in the performance attainable by standard learning methods which assume a balanced class distribution. Plenty of works [11, 12, 13, 14, 15, 16, 17, 18] have shown the importance of this problem for classification, and lots of studies [19, 20, 21, 22, 23, 24, 25, 26, 27] demonstrated that sample balancing provides a significant quality improvement in real-world application problems including control of financial risk, image recognition, medicine, biology, text mining, time series, etc. Unbalanced data had been identified as one of 10 challenging problems in data mining research long ago [28]. In recent years, classification problem for imbalanced datasets has been a more and more hot research topic.

Sun et al. [29] investigated the nature of imbalanced classification problem and believed that imbalanced class distribution, small sample size, class separability and within-class concepts were the main factors causing the problem. Corresponding to different problem roots, people put forward different schemes of solution to imbalanced classification. For example, in order to tackle the problem of imbalanced class distribution, researchers devised a variety of sampling methods, such as random over- or under-sampling, SMOTE [30] and other methods based on it, CBO [31]. Also a lot of pertinent algorithms (e.g. cost sensitive learning, one-class model, and so on) are developed to alleviate the problem of class imbalance in classification. The details of solution to problems of imbalanced classification can be found in Section 2. Of all these solutions, SMOTE oversampling has received much more attention in recent research. After that, researchers proposed some other methods based on it, including Borderline-SMOTE [32], ADASYN [33], SMOTEboost [34], CBSO [35], RAMOBoost [36], KSMOTE [37]. Their experiments show that all these methods can alleviate the imbalance problem more or less. Despite the modifications in various scopes, the key idea of these methods is still essentially analogue to SMOTE. Our investigation in this paper suggests that a severe flaw exists for this type of oversampling. Specifically, new samples created in SMOTE invariably lie in the line segment between seed samples. This is why new synthetic samples cannot really simulate the distribution of original samples. Besides, nearest neighbors searching algorithm used in SMOTE does not take the distribution of subclasses into account, which can lead to the problem of overlapping between classes. We propose a new oversampling method SNOCC to overcome above defect. The basic idea is to employ an improved method to avoid new synthetic samples confining in the line segment between seed samples, and a different nearest neighbors searching algorithm is adopted to tackle the problem of classes overlapping. SNOCC integrate cluster based on distance and oversampling that is different from SMOTE. Our experiments show that SNOCC oversampling can generate better synthetic samples than SMOTE and other methods based on it.

Our paper is organized as follows. Section 2 briefly introduces the advancements in the domain related to imbalanced data. In Section 3 we analyze the SMOTE method and give the reason why it may cause a generalization error. Section 4 describes our SNOCC oversampling methods. Section 5 presents our experiment result and compares it with results of other different methods. Finally, we present our conclusions in the Section 6.

2 RELATED WORKS

Solutions to problems of imbalanced classification fall into two categories: data-level methods and algorithm-level methods. The former try to relax the skewed degree of dataset through the adjustment the distribution of classes, and the latter is to design new classifier or improve the existing algorithms to increase the recognition ratio of the positive class. Typical approaches of algorithm-level methods include cost-sensitive learning [38], one-class learning [39, 40], ensemble learning [41, 42], etc. These types of methods usually involve re-designing of a specific classifier, which is out of the scope of our paper. In this paper, we mainly focus on the data-level methods.

Data-level methods are to sample the dataset and create a balanced data distribution, which include oversampling and undersampling. Oversampling increases the number of minority class instances to balance the distribution of classes. The simplest oversampling is random oversampling, which simply duplicate minority instances. The most severe weakness here is that it adds no new information to the dataset and could cause overfitting of classifiers. The corresponding random undersampling method is to stochastically delete some majority class instances so as to adjust the data distribution. The main shortcoming of random undersampling is that it does not make full use of information from deleted instances. Based on these two random sampling methods mentioned, varieties of heuristic methods for sampling are proposed.

1. Synthetic Minority Over-sampling TEchnique (SMOTE)

   SMOTE [30] is a type of oversampling method. Its theory basis is that the feature space of minority class instances is similar. For each instance x_i in minority class, SMOTE searches its k nearest neighbors and one neighbor is randomly selected as x' (we call instances x_i and x' seed sample). Then a random number between $[0, 1]$ δ is generated. The new artificial sample x_{new} is created as:

$$x_{new} = x_i + (x' - x_i) \times \delta \tag{1}$$

Comparing with random oversampling method, SMOTE method can effectively avoid the problem of overfitting of classifiers. The proposition of SMOTE greatly inspired researches and some derivations of SMOTE such as Borderline-SMOTE [32], ADASYN [33], SMOTEboost [34], CBSO [35], RAMOBoost [36], KSMOTE [37], were put forward. For more detailed analysis of SMOTE, see Section 3.

2. Sampling methods based on data cleaning

   Data cleaning technique can be used to clear out overlapped samples introduced in oversampling. NCL (Neighborhood Clean Rule) [43] is based on improved ENN (Edited Nearest Neighbor). ENN removes a sample where the class label is different from others among three nearest neighbors. NCL method searches 3 nearest neighbors of sample x. If x belongs to the majority class, and at least two of 3 nearest neighbors belong to the minority class, then x should be cleared out. If x belongs to the minority class, and more than two of 3 nearest neighbors belong to the majority class, then the majority class instances of 3 nearest neighbors should be deleted. Tomek links are also often-used as a data cleaning method. Tomek links method is defined as a pair of samples that belong to different classes and the distance to each other is the nearest. We can delete Tomek links to clear samples generated from oversampling until all nearest neighbors pairs belong to the same class.

3. Cluster-Based Oversampling (CBO)

   CBO [31] is proposed to effectively manage the within-class imbalance problem. It makes use of the K-means clustering technique. Before oversampling, both minority and majority class samples should be clustered. In the majority class, all the clusters, except for the largest one, are randomly oversampled so as to get comparable number of training samples as the largest cluster. In the minority class, all the clusters are randomly oversampled to get the same number until the sample size of the minority is equal to that of majority.

3 DEFECTS OF SMOTE-BASED APPROACH

In this section we illustrate using examples that SMOTE-based oversampling will cause problems on the distribution of samples and hinder the accuracy of the classifiers. New synthetic samples in SMOTE cannot really reflect the distribution of original samples. When we learn classification model from these new samples, the final classifier cannot obtain correct information on real samples data. This will increase the probability that classifier misclassifies samples and lead to a bigger generalization error.

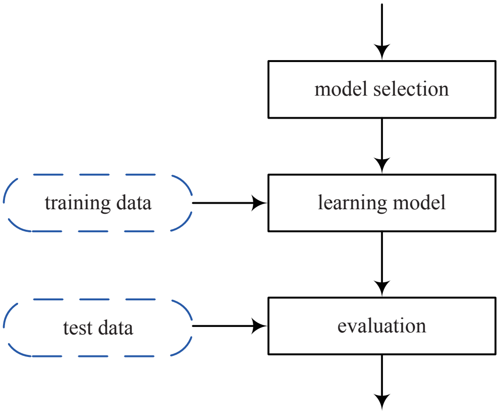

Figure 1. The process of classification

In order to analyze SMOTE methods, we begin with classification. In classification, classifier is used to assign a label to a sample with unknown class label according to a sequence of rules. Such classifier is obtained through learning the training data with known class labels. A standard classification process is shown in Figure 1. First, we need to choose a proper classification method. Then the selected model learns training data to generate final classifier. Finally, test data is used to assess the classifier.

After choosing a model, the performance of a classifier is to a great extent determined by the training-sample size. Features selection methods and the choice of classification algorithm had only a modest impact on the predictor performance [44]. The study of GN Karystinos et al. shows that the generalization error may be undesirably large when the available training set size is too small [45]. In addition, the within-class imbalance phenomena, which correspond to the imbalanced distribution among subclasses [46], can also greatly impair the generalization ability of the classifier. If the distribution of training data cannot include the whole samples space of test data, it is nothing surprising that we will get a classifier with a greater generalization error.

Now let us return to SMOTE. From Equation (1) we can see that the new sample x_{new} is created by the linear interpolation of two seed samples x_i and x'. As a result, the generated synthetic sample x_{new} lies in the line segment between x_i and x'. Figures 2 and 3 show the distribution of new samples generated by SMOTE oversampling. In Figure 2, there are 3 seed samples and 100 new synthetic samples generated by SMOTE. In Figure 3, there are 10 seed samples and 500 new synthetic samples generated by SMOTE. From these two figures, we can intuitively say that the space of new synthetic samples generated by SMOTE cannot cover the distribution area of seed samples. In other word, we may say that new synthetic samples generated by SMOTE cannot correctly reproduce the distribution of original seed samples. When these new samples serve as training data, it is inevitable that this will more or less cause additional error.

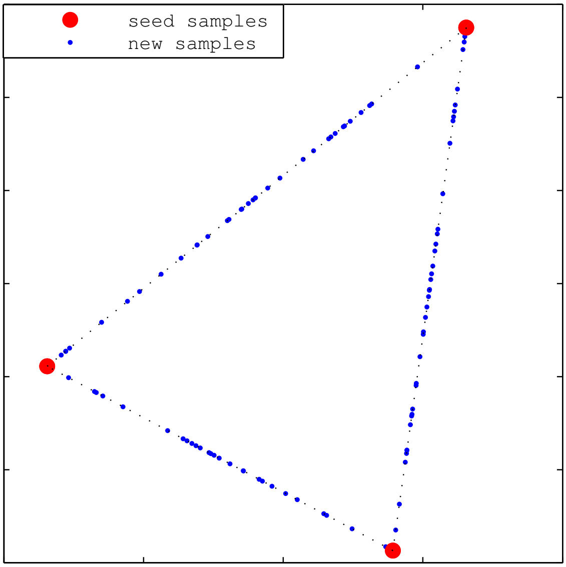

Figure 2. The distribution of new samples for oversampling 3 samples 100 times using SMOTE

We implement an example for classification to validate this statement. Figure 4 shows the distribution of positive and negative samples in original data. In the test, we choose the vertex of polygon, which contains all positive samples, as seed samples of SMOTE. Then positive training samples are created by SMOTE. Negative training samples are randomly selected from negative samples. Training set consists of positive and negative training samples. The rest samples of original data are used as a test set. The result of classification is shown as Figure 5. From Figure 5 we can find that most of misclassified positive samples lie in these triangular areas divided by straight lines. This further confirms that if the test samples lie in the area which the training dataset cannot cover, they will be misclassified in a large probability. Since classifier cannot learn any information about these samples from training data, when they appear in the test data, it is perfectly normal that they will be misclassified. Our test also shows that for different classification models or learning parameters, the probability that these samples are misclassified is different. The reason is that there are differences in generalization abilities of classification models or learning parameters.

Based on the experiments and analysis above we can conclude that the distribution space of new samples created in SMOTE is confined in the line segment between

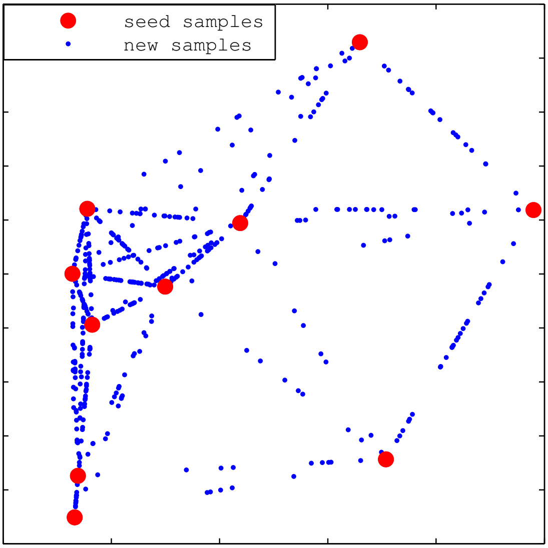

Figure 3. The distribution of new samples for oversampling 10 samples 500 times using SMOTE

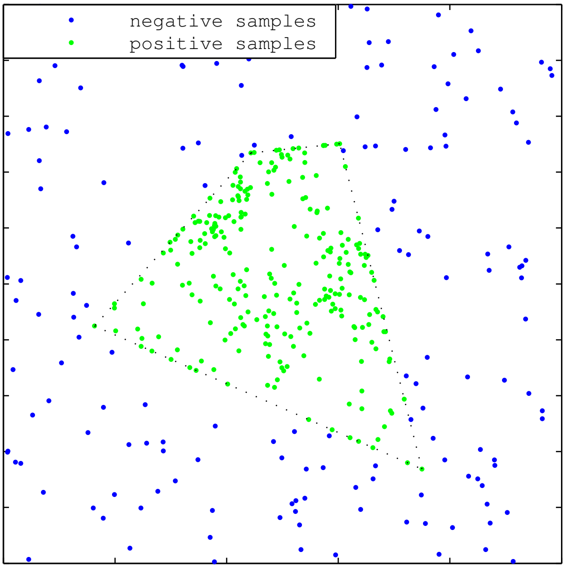

Figure 4. The distribution of all positive and negative samples in example for classification

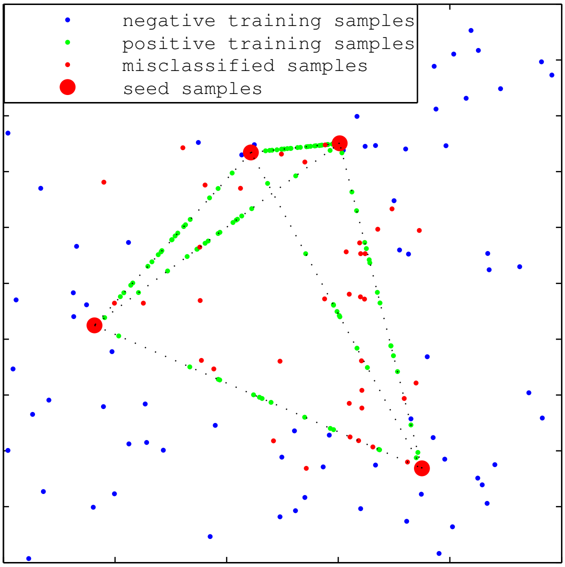

Figure 5. The distribution of training samples and misclassified samples (marked in red) in the example for classification

seed samples and it cannot reproduce the distribution of original samples. This is an important cause resulting in a bad performance of classifier. Since classifier cannot get real information on original sample, there is a bigger probability that samples are misclassified. Besides, SMOTE does not consider neighboring samples and that may cause decision boundaries for the minority class to spread further into the majority class space. This can result in the problem of overlapping between classes [35, 47].

4 SIGMA NEAREST OVERSAMPLING BASED ON CONVEX COMBINATION

In consideration of defects of SMOTE, we propose an oversampling based on convex combination (SNOCC, Sigma Nearest Oversampling based on Convex Combination).

Convex combination is a linear combination of points where all coefficients are non-negative and sum up to 1. Given a finite number of points $x_1, x_2, \ldots, x_n$ in a real vector space, a convex combination of these points is a point of the form:

$$x = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n \tag{2}$$

where the real numbers α_i $(i = 1, 2, \ldots, n)$ satisfy $\alpha_i \geq 0$ and $\alpha_1 + \alpha_2 + \ldots + \alpha_n = 1$.

In SMOTE, according to formula (1) we can get $x_{new} = (1 - \delta) \times x_i + \delta \times x'$. We may tell that x_{new} is the convex combination of x_i and x' according to $(1 - \delta) + \delta = 1$. This is the simplest form of convex combination ($n = 2$). We extend formula (1) in order to new samples can distribute through sample space. In our oversampling method, we generate new samples according to formula (2). The detailed oversampling procedure of SNOCC is described in Algorithm 1.

Algorithm 1 Framework of SNOCC oversampling

```
Input:
    Input Seed samples set of minority class, $S$
    Amount of new samples, $N$
    Number of nearest neighbors, $k$
Output:
    Output $N$ new samples of minority class
1: For every sample $s_i$ in the seed samples of minority class $S$, calculate its $k$ nearest neighbors.
2: Calculate the mean $m_i$ of distance from every sample $s_i$ to its $k$ nearest neighbors. And let *sigma* equal to the average $m_i$ plus the standard deviation *std* of all $m_i$.
3: For every sample $s_i$ in $S$, search all nearest neighbors that their distance to $s_i$ is not greater than *sigma* and mark their corresponding index. We call them sigma nearest neighbors.
4: Randomly select one sample $s_1$ from $S$, and then randomly select two nearest neighbors ($s_2$ and $s_3$) of $s_1$.
5: Generate a random non-negative vector of 3 dimensions $\alpha$ ($\alpha_1$, $\alpha_2$, $\alpha_3$) and normalize $\alpha$ so that $\alpha_1 + \alpha_2 + \alpha_3 = 1$.
6: Take sample $s_1$ and its two sigma nearest neighbors ($s_2$ and $s_3$) as seed samples, generate one new sample $s$ according to formulation $s = \alpha_1 s_1 + \alpha_2 s_2 + \alpha_3 s_3$.
7: Repeat step 4, 5, 6 $N$ times.
8: return $N$ new samples
```

In SNOCC, we refer $x_1, x_2, \ldots, x_n$ as seed samples. In the course of oversampling, the number of seeds (n) can be adjusted. For different n, the corresponding distribution domains are different. Figures 6 and 2 show the distribution maps of 100 new samples of SNOCC and SMOTE oversampling for 3 original samples respectively. It is obvious that the distribution domain of new samples is confined on the line segments where endpoints are the seed samples for SMOTE oversampling (Figure 2). For SNOCC oversampling (Figure 6), the new samples can be any position of a convex hull constituted of seed samples.

Figures 7 and 3 are the distributions of new samples for oversampling 10 samples 500 times using SNOCC and SMOTE respectively. Figures 8 and 9 are the distributions of new samples for oversampling 20 samples 500 times using SNOCC and SMOTE respectively. It can be seen in Figures 5 and 11 that the distribution domain of new samples in SMOTE oversampling is located at the line segments of

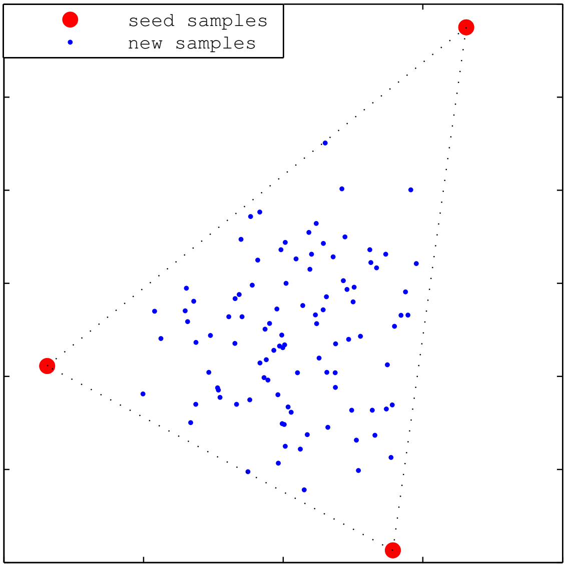

Figure 6. The distribution of new samples for oversampling 3 samples 100 times using SNOCC

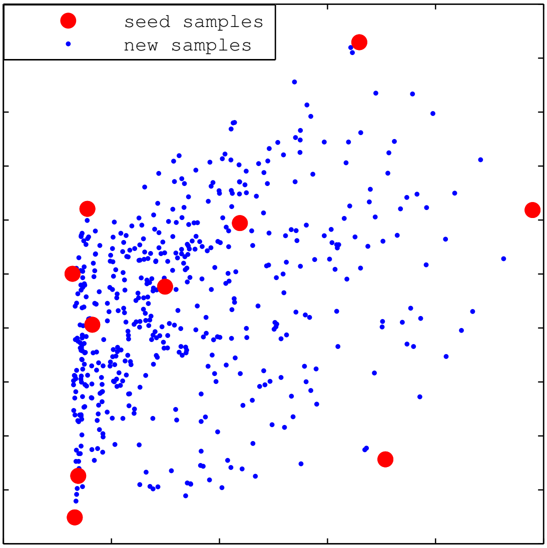

Figure 7. The distribution of new samples for oversampling 10 samples 500 times using SNOCC

original samples. Comparing to Figures 7 and 3, Figures 8 and 9 show that new samples of SNOCC oversampling can better reproduce the distribution of original samples, even if the distribution is irregular (Figure 8).

It is natural for SNOCC to handle continuous feature. For ordinal feature, we first map the values into an integer's sequence. For example, if there are n distinct values in an ordinal feature. The integer's sequence is $(1, 2, \ldots, n)$. During oversampling, it is not the original value in the ordinal feature but the integer's sequence calculated. The corresponding results are rounded to the nearest integer. Finally, this integer will be mapped back into original value in the ordinal feature according to the inverse mapping used above.

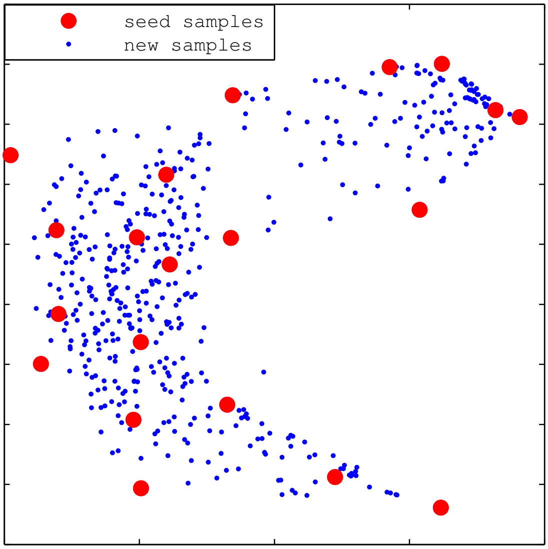

Figure 8. The distribution of new samples for oversampling 20 samples 500 times using SNOCC

5 EXPERIMENTS

We conducted experiments to validate the efficiency of SNOCC. In order to compare the performance of SNOCC and SMOTE oversampling, we implemented SMOTE and CBSO [35], which is a new derivation of SMOTE. In our experiments, all twelve datasets are from KEEL-dataset [48]. All datasets' IR (*Imbalance Ratio*, the asymmetry degree of imbalanced data, which is the ratio of the number of negative examples to that of positive examples) is greater than 9. Information on datasets is shown in Table 1.

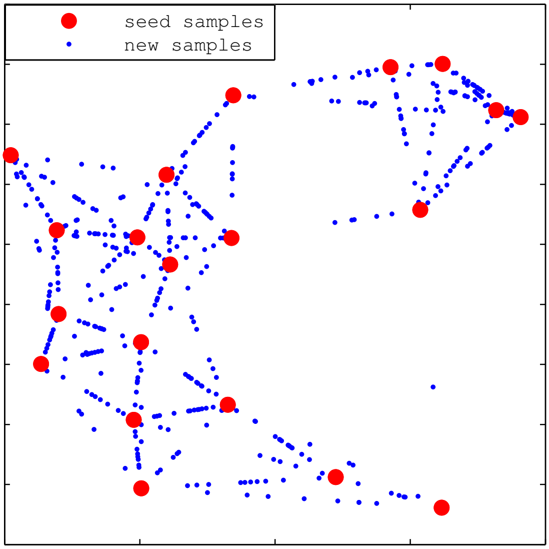

Figure 9. The distribution of new samples for oversampling 20 samples 500 times using SMOTE

We employed naive-Bayes classifier to evaluate the efficiency of SMOTE, CBSO and SNOCC. Variables are discretized before applying the classifier. All methods used the same scheme of discretization. We used Laplace estimate to calculate the prior probability. Laplace estimate shows good performance in naive-Bayes classification [49]. One extra benefit of using Laplace estimate is that zero probability can be avoided. F-measure for the minority (positive) class was used as the assessment standard.

In our experiments, we used 3-fold cross-validation to measure the performance of classifier learned from training dataset which is generated through different oversampling methods. The reason we do not adopt more folds but 3 folds is that the number of positive samples is less than 10 in several datasets. In 3-fold cross-validation, the whole dataset is randomly partitioned into 3 folds and each fold contains approximately the same proportion of classes as the original datasets. Of the 3 folds, a single fold is retained as the validation data for the model testing, and the remaining 2 folds are used as training data. The cross-validation process is then repeated 3 times, with each of the 3 folds used exactly once as the validation data. The 3 results from the folds then are averaged to produce the estimation of one test.

Our experiment process is shown in Figure 10, which consists of 6 steps. All steps are described below:

Step 1: The whole imbalanced dataset is randomly divided into training dataset and test dataset.

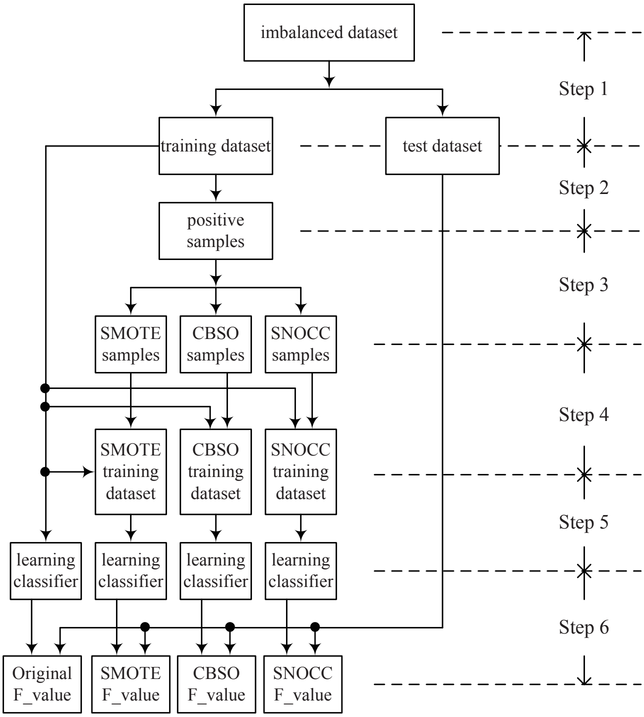

Figure 10. The flowchart of experiment

Dataset	Number of Features	Sample Size	Number of Positive Samples	Number of Negative Samples	Imbalance Ratio
ecoli-0-1-3-7_vs_2-6	7	281	7	274	39.14
ecoli4	7	336	20	316	15.8
glass-0-1-6_vs_5	9	184	9	175	19.44
glass5	9	214	9	205	22.78
yeast-0-5-6-7-9_vs_4	8	528	51	477	9.35
yeast-1-2-8-9_vs_7	8	947	30	917	30.57
yeast-1-4-5-8_vs_7	8	693	30	663	22.1
yeast-1_vs_7	7	459	30	429	14.3
yeast-2_vs_4	8	514	51	463	9.08
yeast4	8	1484	51	1433	28.1
yeast5	8	1484	44	1440	32.73
yeast6	8	1484	35	1449	41.4

Table 1. Information of datasets

Step 2: All positive samples in training dataset are picked up, and the numbers of positive and negative samples are counted. The number of new positive samples to be generated is the difference of the number of negative samples and that of positive samples.

Step 3: Oversampling methods SMOTE, CBSO and SNOCC are called to create new synthetic SMOTE, CBSO and SNOCC samples, respectively.

Step 4: Adding SMOTE, CBSO and SNOCC samples into original training dataset to form new SMOTE, CBSO and SNOCC training dataset.

Step 5: Learning classifier from original, SMOTE, CBSO and SNOCC training dataset, respectively.

Step 6: Classifying the samples in test dataset using classifier learned from Step 5 and calculating corresponding F-measure *F-value* for positive class.

In order to make a fair comparison with SMOTE, the number of nearest neighbors k in SNOCC was set to 5, which is the same as the default value in SMOTE. After oversampling, we obtained a training dataset of positive samples and negative samples in equal portions, that is, a balanced dataset. To eliminate random effects, for each dataset, we ran each oversampling algorithms 100 times and call naive-Bayes classifier to get corresponding *F-value* for each time, resulting in 100 *F-value* for each oversampling algorithm. Finally, t-test was used to verify the significance of the *F-value* differences between methods.

We computed the mean values and standard deviations of the *F-value* of 100 classifications without oversampling, with SMOTE oversampling, with CBSO oversampling and with SNOCC oversampling, respectively. The results are shown in Table 2. In each dataset, the biggest among all *F-value* pertained to different oversampling methods is bold-faced. Of all twelve biggest *F-value*, SNOCC accounts for nine and the rest three are from CBSO. This shows that in generating new synthetic samples, SNOCC performs much better. t-test was performed to compare the significance of the results obtained from SNOCC with SMOTE, and SNOCC with CBSO, respectively, at 0.05 significance level. The test results are presented in Table 3. By combining with Table 2, we can determine which one is the winner. The h is 1^+ if SNOCC is winner and 1^- otherwise. From Table 3 we can find that SNOCC outperforms SMOTE on ten of twelve datasets. And SMOTE outperforms SNOCC on two datasets. SNOCC outperforms CBSO on eight of twelve datasets and there are three datasets that CBSO outperforms SNOCC. On the whole, the experiments results show that the SNOCC performances are significantly better than those of SMOTE and CBSO.

6 CONCLUSIONS

After decades of development, classification techniques get matured day by day. But most of the existing classifiers tend to identify majority class samples and usually fail to classify minority class samples with a satisfactory accuracy. There are plenty

Dataset	RAW	SMOTE	CBSO	SNOCC
ecoli-0-1-3-7_vs_2-6	0.2992±0.1906	0.5100±0.1404	0.5028±0.1336	**0.5749±0.1413**
ecoli4	0.3626±0.0746	0.6723±0.0539	**0.6850±0.0586**	0.6552±0.0484
glass-0-1-6_vs_5	0.5597±0.1651	0.5620±0.0445	0.6186±0.0685	**0.7219±0.0855**
glass5	0.5247±0.1394	0.5407±0.0402	0.5801±0.0743	**0.6651±0.0871**
yeast-0-5-6-7-9_vs_4	0.0356±0.0258	0.3272±0.0212	**0.3459±0.0262**	0.3363±0.0246
yeast-1-2-8-9_vs_7	0.0265±0.0348	0.0830±0.0048	0.0826±0.0044	**0.0947±0.0116**
yeast-1-4-5-8_vs_7	0.0091±0.0216	0.1100±0.0051	0.1079±0.0056	**0.1286±0.0138**
yeast-1_vs_7	0.1311±0.0658	0.2463±0.0235	0.2365±0.0183	**0.2613±0.0380**
yeast-2_vs_4	0.6057±0.0419	0.6739±0.0268	**0.6748±0.0301**	0.6630±0.0320
yeast4	0.0457±0.0289	0.1620±0.0094	0.1813±0.0108	**0.1818±0.0114**
yeast5	0.4758±0.0540	0.5279±0.0125	0.5368±0.0121	**0.5747±0.0201**
yeast6	0.1882±0.0703	0.2041±0.0141	0.2288±0.0169	**0.3903±0.0309**

This table shows F-measure value of classification pertained to different oversampling methods. Oversampling methods is the column title. The second column titled RAW is the *F-value* without oversampling. Each grid is filled with average *F-value±standard derivation*. In each row, the biggest *F-value* is bold-faced.

Table 2. F-values of different oversampling methods

of imbalanced data in the application domain. The basic reason is that either it is very difficult to collect data or positive samples in collected data are rare [50, 51]. This poses a challenge to academic community. SMOTE oversampling proposed by Chawla et al. gave us a good start to tackle the problem of imbalanced distribution. And based on SMOTE, researchers did a lot of fruitful work.

Even so, we find some weakness of SMOTE. In classification, the distribution of training data can greatly influence the generalization ability of a classifier. New samples generated by SMOTE oversampling are confined in the line segment between two seed samples (Figures 2, 3 and 9). That means that new samples created by SMOTE oversampling cannot fully cover the distribution space of original samples. This is a major factor that causes generalization error of the classifier. Besides, SMOTE introduces the problem of classes overlapping [35, 47].

SNOCC proposed in the paper can remedy the defect of SMOTE. In SNOCC, the method that creates new samples makes each new sample likely to locate in any place of convex hull formed by seed samples. We can see this from Figures 6, 7 and 8. In order to minimize the adverse effects of classes overlapping on the performance of classification, we do not use k-nearest-neighbors but nearest-neighbors based on distance to search neighbors of samples. This change can improve the efficiency of oversampling and our results of experiment also support this conclusion. SNOCC method can generate samples that naturally model the distribution of original samples. Our experiment results show that SNOCC outperforms SMOTE and methods derived from it. Our method failed in 2 and 3 datasets, respectively. However, the difference of performance in these datasets between our method and the others is marginal. We can see that from Table 3. The smallest p value was 0.000131774 in these datasets while it was less than 10^{-8} in most datasets that our method won. We think that the possible reason which causes this difference is

dataset	**t-test 1**		**t-test 2**	
	h	**p**	**h**	**p**
ecoli-0-1-3-7_vs_2-6	1^+	0.001391304	1^+	0.00029264
ecoli4	1^-	0.019986531	1^-	0.000131774
glass-0-1-6_vs_5	1^+	4.63×10^{-39}	1^+	1.53×10^{-17}
glass5	1^+	4.51×10^{-28}	1^+	3.97×10^{-12}
yeast-0-5-6-7-9_vs_4	1^+	0.006099216	1^-	0.008741365
yeast-1-2-8-9_vs_7	1^+	2.25×10^{-17}	1^+	1.89×10^{-18}
yeast-1-4-5-8_vs_7	1^+	2.99×10^{-27}	1^+	6.01×10^{-31}
yeast-1_vs_7	1^+	0.000997247	1^+	1.89×10^{-8}
yeast-2_vs_4	1^-	0.010792733	1^-	0.008674209
yeast4	1^+	3.57×10^{-29}	0	0.777325853
yeast5	1^+	2.51×10^{-48}	1^+	9.15×10^{-38}
yeast6	1^+	4.03×10^{-121}	1^+	7.13×10^{-107}

Column t-test 1 shows the result of t-test that tests the significant difference between *F-value* of SMOTE and that of SNOCC with the 5 % significance level. And sub-column h is the result and p the corresponding p value. A 1^+ indicates that the performance of SNOCC method is significantly better than that of SMOTE, and 1^- indicates a reverse result.

Column t-test 2 shows the result of t-test that tests the significant difference between *F-value* of CBSO and that of SNOCC with the 5 % significance level. And sub-column h is the result and p the corresponding p value. A 1^+ indicates that the performance of SNOCC method is significantly better than that of CBSO, and 1^- indicates a reverse result.

The zero value of h shows that there is no significant difference between two methods at 5 % significance level.

Table 3. The result of t-test

that the problem of overlapping between classes in these data sets either is not so bad, or does not exist at all. So, there are few or no negative effects on SMOTE or CBSO. Besides, comparing it to these benchmark methods, from the result of our experiment we can see that our method can obtain much better performance in data sets with greater imbalance ratio. More experiments will be done to improve it in the next work. In our work, SNOCC can only handle continuous and ordinal feature. Future work will be focused on how to deal with categorical and Boolean attribute.

Acknowledgements

This work was supported in part by Shenzhen Innovation Funding for Advance Talents (No. KQCX20130628112914291), the Promotion and Development Project for Key Laboratory of Shenzhen (No. CXB201104220026A), the National Natural Science Foundation of Youth Science Foundation (No. 31000447), the National Natural Science Foundation of China (No. 61363005, 61462017, 61462018), Guangxi Natural Science Foundation (No. 2014GXNSFDA118036, 2013GXNSFAA019324), Education Department of Guangxi (No. ZD2014049, 2013YB084), Guangxi Key Laboratory of Trusted Software (No. kx201512). The authors would also like to thank the anonymous reviewers for their helpful comments and suggestions for improving the manuscript.

REFERENCES

[1] Zhang, J.—Bloedorn, E.—Rosen, L.—Venese, D.: Learning Rules from Highly Unbalanced Data Sets. Fourth IEEE International Conference on Data Mining (ICDM '04), IEEE, 2004, pp. 571–574.

[2] Phua, C.—Alahakoon, D.—Lee, V.: Minority Report in Fraud Detection: Classification of Skewed Data. ACM SIGKDD Explorations Newsletter, Vol. 6, 2004, No. 1, pp. 50–59.

[3] Sales, A. P.—Tomaras, G. D.—Kepler, T. B.: Improving Peptide-MHC Class I Binding Prediction for Unbalanced Datasets. BMC Bioinformatics, Vol. 9, 2008, Art. No. 385.

[4] Chen, X.—Jeong, J. C.: Sequence-Based Prediction of Protein Interaction Sites with an Integrative Method. Bioinformatics, Vol. 25, 2009, No. 5, pp. 585–591.

[5] Batuwita, R.—Palade, V.: microPred: Effective Classification of Pre-miRNAs for Human miRNA Gene Prediction. Bioinformatics, Vol. 25, 2009, No. 8, pp. 989–995.

[6] Wu, J.—Liu, H.—Duan, X.—Ding, Y.—Wu, H.—Bai, Y.—Sun, X.: Prediction of DNA-Binding Residues in Proteins from Amino Acid Sequences Using a Random Forest Model with a Hybrid Feature. Bioinformatics, Vol. 25, 2009, No. 1, pp. 30–35.

[7] Forman, G.: An Extensive Empirical Study of Feature Selection Metrics for Text Classification. Journal of Machine Learning Research, Vol. 3, 2003, pp. 1289–1305.

[8] Mladenic, D.—Grobelnik, M.: Feature Selection for Unbalanced Class Distribution and Naive Bayes. Proceedings of the Sixteenth International Conference on Machine Learning (ICML '99), 1999, pp. 258–267.

[9] Mazurowski, M. A.—Habas, P. A.—Zurada, J. M.—Lo, J. Y.—Baker, J. A.—Tourassi, G. D.: Training Neural Network Classifiers for Medical Decision Making: The Effects of Imbalanced Datasets on Classification Performance. Neural Networks, Vol. 21, 2008, No. 2-3, pp. 427–436.

[10] Cohen, G.—Hilario, M.—Sax, H.—Hugonnet, S.—Geissbuhler, A.: Learning from Imbalanced Data in Surveillance of Nosocomial Infection. Artificial Intelligence in Medicine, Vol. 37, 2006, No. 1, pp. 7–18.

[11] Lara, L.—Blagus, R.: SMOTE for High-Dimensional Class-Imbalanced Data. BMC Bioinformatics, Vol. 14, 2013, Art. No. 106.

[12] García, V.—Sánchez, J. S.—Mollineda, R. A.: On the Effectiveness of Preprocessing Methods when Dealing with Different Levels of Class Imbalance. Knowledge-Based Systems, Vol. 25, 2012, No. 1, pp. 13–21.

[13] Barua, S.—Islam, M.—Yao, X.—Murase, K.: MWMOTE – Majority Weighted Minority Oversampling Technique for Imbalanced Data set Learning. IEEE Transactions on Knowledge and Data Engineering, Vol. 26, 2014, No. 2, pp. 405–425.

[14] Menardi, G.—Torelli, N.: Training and Assessing Classification Rules with Imbalanced Data. Data Mining and Knowledge Discovery, Vol. 28, 2014, No. 1, pp. 92–122.

[15] Galar, M.—Fernández, A.—Barrenechea, E.—Herrera, F.: EUSBoost: Enhancing Ensembles for Highly Imbalanced Data-Sets by Evolutionary Undersampling. Pattern Recognition, Vol. 46, 2013, No. 12, pp. 3460–3471.

[16] Anand, A.—Pugalenthi, G.—Fogel, G. B.—Suganthan, P. N.: An Approach for Classification of Highly Imbalanced Data Using Weighting and Undersampling. Amino Acids, Vol. 39, 2010, No. 5, pp. 1385–1391.

[17] López, V.—Triguero, I.—Carmona, C. J.—García, S.—Herrera, F.: Addressing Imbalanced Classification with Instance Generation Techniques: IPADE-ID. Neurocomputing, Vol. 126, 2014, pp. 15–28.

[18] Ramentol, E.—Caballero, Y.—Bello, R.—Herrera, F.: SMOTE-RSB*: A Hybrid Preprocessing Approach Based on Oversampling and Undersampling for High Imbalanced Data-Sets Using SMOTE and Rough Sets Theory. Knowledge and Information Systems, Vol. 33, 2012, No. 2, pp. 245–265.

[19] Menardi, G.—Torelli, N.: Effect of Training Set Selection when Predicting Defaulter SMEs with Unbalanced Data. Working Paper Series, No. 1, 2010, DEAMS, Universita degli Studi di Trieste. 13 p.

[20] Song, Y.—Morency, L.-P.—Davis, R.: Distribution-Sensitive Learning for Imbalanced Datasets. 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013, pp. 1–6.

[21] Mostafizur Rahman, M.—Davis, D. N.: Cluster Based Under-Sampling for Unbalanced Cardiovascular Data. Proceedings of the World Congress on Engineering, Vol. 3, 2013.

[22] Newby, D.—Freitas, A. A.—Ghafourian, T.: Coping with Unbalanced Class Data Sets in Oral Absorption Models. Journal of Chemical Information and Modeling, Vol. 53, 2013, No. 2, pp. 461–474.

[23] Wu, Q.—Liu, H.—Liu, K.: Mixed-Sampling Approach to Unbalanced Data Distributions: A Case Study Involving Leukemia's Document Profiling. WSEAS Transactions on Information Science and Applications, Vol. 8, 2011, No. 9, pp. 356–379.

[24] Yu, H.—Ni, J.—Zhao, J.: ACOSampling: An Ant Colony Optimization-Based Undersampling Method for Classifying Imbalanced DNA Microarray Data. Neurocomputing, Vol. 101, 2013, pp. 309–318.

[25] Li, Q.—Wang, Y.—Bryant, S. H.: A Novel Method for Mining Highly Imbalanced High-Throughput Screening Data in PubChem. Bioinformatics, Vol. 25, 2009, No. 24, pp. 3310–3316.

[26] Wang, S.—Li, D.—Zhao, L.—Zhang, J.: Sample Cutting Method for Imbalanced Text Sentiment Classification Based on BRC. Knowledge-Based Systems, Vol. 37, 2013, pp. 451–461.

[27] Cao, H.—Li, X.-L.—Woon, D. Y.-K.—Ng, S.-K.: Integrated Oversampling for Imbalanced Time Series Classification. IEEE Transactions on Knowledge and Data Engineering, Vol. 25, 2013, No. 12, pp. 2809–2822.

[28] Yang, Q.—Wu, X.: 10 Challenging Problems in Data Mining Research. International Journal of Information Technology & Decision Making, Vol. 5, 2006, No. 4, pp. 597–604.

[29] SUN, Y.—WONG, A. K. C.—KAMEL, M. S.: Classification of Imbalanced Data: A Review. International Journal of Pattern Recognition and Artificial Intelligence, Vol. 23, 2009, No. 04, pp. 687–719.

[30] CHAWLA, N. V.—BOWYER, K. W.—HALL, L. O.—KEGELMEYER, W. P.: SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research, Vol. 16, 2002, pp. 321–357.

[31] JO, T.—JAPKOWICZ, N.: Class Imbalances Versus Small Disjuncts. ACM SIGKDD Explorations Newsletter, Vol. 6, 2004, No. 1, pp. 40–49.

[32] HAN, H.—WANG, W.-Y.—MAO, B.-H.: Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Advances in Intelligent Computing, LNCS, Vol. 3644, 2005, pp. 878–887.

[33] HE, H.—BAI, Y.—GARCIA, E. A.—LI, S.: ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. IEEE International Joint Conference on Neural Networks (IJCNN 2008), IEEE World Congress on Computational Intelligence, 2008, pp. 1322–1328.

[34] CHAWLA, N.—LAZAREVIC, A.—HALL, L.—BOWYER, K.: SMOTEBoost: Improving Prediction of the Minority Class in Boosting. Knowledge Discovery in Databases: PKDD 2003, LNCS, Vol. 2838, 2003, pp. 107–119.

[35] BARUA, S.—ISLAM, M.—MURASE, K.: A Novel Synthetic Minority Oversampling Technique for Imbalanced Data Set Learning. Neural Information Processing, Springer, LNCS, Vol. 7063, 2011, pp. 735–744.

[36] CHEN, S.—HE, H.—GARCIA, E. A.: RAMOBoost: Ranked Minority Oversampling in Boosting. IEEE Transactions on Neural Networks, Vol. 21, 2010, No. 10, pp. 1624–1642.

[37] PRACHUABSUPAKIJ, W.—SOONTHORNPHISAJ, N.: Clustering and Combined Sampling Approaches for Multi-Class Imbalanced Data Classification. Advances in Information Technology and Industry Applications, 2012, pp. 717–724.

[38] ELKAN, C.: The Foundations of Cost-Sensitive Learning. Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI '01), 2001, Vol. 2, pp. 973–978.

[39] RASKUTTI, B.—KOWALCZYK, A.: Extreme Re-Balancing for SVMs: A Case Study. ACM SIGKDD Explorations Newsletter, Vol. 6, 2004, No. 1, pp. 60–69.

[40] ZHUANG, L.—DAI, H.: Parameter Estimation of One-Class SVM on Imbalance Text Classification. Advances in Artificial Intelligence, LNCS, Vol. 4013, 2006, pp. 538–549.

[41] YAN, R.—LIU, Y.—JIN, R.—HAUPTMANN, A.: On Predicting Rare Classes with SVM Ensembles in Scene Classification. Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03), 2003, Vol. 3, pp. 21–24.

[42] BHATNAGAR, V.—BHARDWAJ, M.—MAHABAL, A.: Comparing SVM Ensembles for Imbalanced Datasets. 2010 IEEE 10th International Conference on Intelligent Systems Design and Applications (ISDA), 2010, pp. 651–657.

[43] LAURIKKALA, J.: Improving Identification of Difficult Small Classes by Balancing Class Distribution. Artificial Intelligence in Medicine, LNCS, Vol. 2101, 2001, pp. 63–66.

[44] Popovici, V.—Chen, W.—Gallas, B. G.—Hatzis, C.—Shi, W.—Samuelson, F. W.—Nikolsky, Y.—Tsyganova, M.—Ishkin, A.—Nikolskaya, T. et al.: Effect of Training-Sample Size and Classification Difficulty on the Accuracy of Genomic Predictors. Breast Cancer Research, Vol. 12, 2010, No. 1, R5.

[45] Karystinos, G. N.—Pados, D. A.: On Overfitting, Generalization, and Randomly Expanded Training Sets. IEEE Transactions on Neural Networks, Vol. 11, 2000, No. 5, pp. 1050–1057.

[46] Japkowicz, N.: Concept-Learning in the Presence of Between-Class and Within-Class Imbalances. Advances in Artificial Intelligence, LNAI, Vol. 2056, 2001, pp. 67–77.

[47] Kotsiantis, S.—Kanellopoulos, D.—Pintelas, P.: Handling Imbalanced Datasets: A Review. GESTS International Transactions on Computer Science and Engineering, Vol. 30, 2006, No. 1, pp. 25–36.

[48] Keel-dataset repository: `http://keel.es/datasets.php`.

[49] Provost, F.—Domingos, P.: Tree Induction for Probability-Based Ranking. Machine Learning, Vol. 52, 2003, No. 3, pp. 199–215.

[50] Chawla, N. V.—Japkowicz, N.—Kotcz, A.: Editorial: Special Issue on Learning from Imbalanced Data Sets. ACM SIGKDD Explorations Newsletter, Vol. 6, 2004, No. 1, pp. 1–6.

[51] Japkowicz, N.—Stephen, S.: The Class Imbalance Problem: A Systematic Study. Intelligent Data Analysis, Vol. 6, 2002, No. 5, pp. 429–449.

Zhuoyuan Zheng received his B.Sc. and M.Sc. degrees in engineering in computer application technology from Guilin University of Electronic Technology, doctorate of computer application technology from University of Chinese Academy of Sciences, P.R. China in 2000 and 2003, respectively, and his Ph.D. degree in computer application technology from University of Chinese Academy of Sciences, P.R. China in 2015. Now, he is a lecturer in Guilin University of Electronic Technology. His current research interests include machine learning, pattern recognization and image processing.

Yunpeng Cai received his Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2007. He is currently Associate Professor with the Center for Biomedical Information Technology, Shenzhen Institute of Advanced Technology (SIAT), Chinese Academy of Sciences, Shenzhen, China. His research interests include bioinformatics, health informatics, machine learning and evolutionary computation.

Ye Li received his B.Sc. and M.Sc. degrees in electrical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 1999 and 2002, respectively, and his Ph.D. degree in electrical engineering from the Arizona State University, Tempe, U.S., in 2006. In 2007, he worked in Cadence Design Systems Inc., San Jose, CA. Since 2008, he has been working as Assistant Professor in the Institute of Biomedical and Health Engineering, Shenzhen Institute of Advanced Technology (SIAT), Chinese Academy of Sciences, Shenzhen, China. He is also the Executive Director of the Center for Biomedical Information Technology. His research interests include low-power wireless system design, sensor communication protocol design, and low power digital IC design.