Chem-Bio Informatics Journal, Vol.13, pp.19-29 (2013)

A Novel Over-Sampling Method and its Application to Cancer Classification from Gene Expression Data

Xuan Tho Dang[1*], Osamu Hirose[2], Duong Hung Bui[3], Thammakorn Saethang[1],
Vu Anh Tran[1], Lan Anh T. Nguyen[1], Tu Kien T. Le[1], Mamoru Kubo[2], Yoichi Yamada[2],
Kenji Satou[2]

[1]Graduate School of Natural Science and Technology, Kanazawa University, Kakuma-machi, Kanazawa 920-1192, Japan
[2]Institute of Science and Engineering, Kanazawa University, Kakuma-machi, Kanazawa 920-1192, Japan
[3]Faculty of Information Technology, Vietnam Trade Union University, 169 Tay Son Road, Hanoi, Vietnam

**E-mail: thodx@hnue.edu.vn*

(Received November 16, 2012; accepted January 4, 2013; published online January 30, 2013)

Abstract

One of the most critical and frequent problems in biomedical data classification is imbalanced class distribution, where samples from the majority class significantly outnumber the minority class. SMOTE is a well-known general over-sampling method used to address this problem; however, in some cases it cannot improve or even reduces classification performance. To address these issues, we have developed a novel minority over-sampling method named safe-SMOTE. Experimental results from two gene expression datasets for cancer classification (i.e., colon-cancer and leukemia) and six imbalanced benchmark datasets from the UCI Machine Learning Repository showed that our method achieved better sensitivity and G-mean values than both the control method (i.e., no over-sampling) and SMOTE. For example, in the colon-cancer dataset, although the sensitivity and specificity achieved by SMOTE (81.36% and 88.63%) were lower than for the control method (81.59% and 89.50%), safe-SMOTE in contrast had these values increase (81.82% and 90.50%). Similarly, the G-mean value of the control (85.45%) decreased to 84.91% when SMOTE was employed, but increased to 86.04% when using safe-SMOTE. In the leukemia dataset, SMOTE was able to improve the sensitivity and G-mean values with respect to the control; however, safe-SMOTE achieved noticeable, even greater improvements for both of these criteria.

Key Words: Imbalanced dataset, SMOTE, Over-sampling, Cancer classification

Area of Interest: Bioinformatics and Bio Computing

1. Introduction

One of the most critical and frequent problems in biomedical data classification is imbalanced class distribution, where samples from the majority class significantly outnumber those from the minority class. The main problem with class imbalances is that typical machine learning methods are often biased to the majority class. As a result, the majority class samples are well classified, whereas many samples from the minority class are easily misclassified. In recent years, the number of imbalanced biomedical datasets has increased, such as microRNA (miRNA) gene prediction [1], protein network analysis [2], and detection of non-coding RNA [3].

Batuwita et al. [1] developed an effective system to classify human precursor microRNA (pre-miRNAs) hairpins from both genome pseudo hairpins and other non-coding RNAs (ncRNAs). In their study, the experimental datasets included three kinds of non-redundant human sequences: 691 pre-miRNAs (positive), 8,494 pseudo hairpins (negative), and 754 other ncRNAs (negative; 9,248 hairpins in total). The class imbalance (i.e. positive-to-negative) ratio of the dataset was determined to be 1:13.4. Radivojac et al. [2] considered designing a complete classification system in protein databases to understand, in detail, protein function and associated complex networks of interactions with other molecules in biochemical pathways. Some of the common characteristics of protein datasets uncovered in this study were that they are often noisy, high-dimensional, sparse, and have class imbalance. The research resulted in the construction of six datasets: e.g., PHOSS (613 positive samples and 10,798 negative samples), PHOST (140 positive samples and 9,051 negative samples), CAM (942 positive samples and 17,974 negative samples), etc. Yu et al. [3] then presented a model for protein-protein interactions (PPIs) that has since aided understanding of the important principles of biological systems. Using the primary structure of proteins, the PPI predictor they developed processes the imbalanced datasets with a positive-to-negative ratio of up to 1:15.

The issue of class imbalance in classification has attracted much research and resulted in a range of publications in the bioinformatics and data mining communities. There are two main strategies to deal with imbalanced class distribution: methods at the data level and methods at the algorithm level. The latter methods aim at adjusting an appropriate inductive bias. For example, Joshi et al. [4] developed a traditional boosting algorithm, which was a promising meta-technique for improving the classification performance of any weak classifiers. With each learning cycle, the boosting algorithm updated the weights of the samples. The weights of the incorrectly classified samples were increased and the weights of the correctly classified samples were decreased. In the following cycles, the classifier focused more on the incorrectly classified samples to achieve higher predictive ability for minority samples. Lin et al. [5] found that the standard Support Vector Machine (SVM) is not suitable for non-standard situations embodied by imbalanced datasets, and proposed a simple procedure for adapting the SVM methodology: using different penalty costs for different classes. Wu et al. [6] also showed that SVMs could be ineffective in determining the class boundary when dealing with imbalanced training-data problems. In order to solve this problem, they proposed to adjust the class boundary based on the kernel function and kernel matrix of SVMs. Based on this analysis, the class-boundary-alignment algorithm worked effectively for imbalanced problems posed by images and video sequences.

There are two data sampling strategies to address class imbalance: over-sampling and under-sampling. In over-sampling, the samples in the minority class are increased, while in under-sampling, the samples in the majority class are decreased; both strategies aim to achieve balanced class distributions as a result. Chawla et al. [7] developed the Synthetic Minority Oversampling Technique (SMOTE), which over-samples minority class samples by generating synthetic minority samples along the line between the minority sample and its nearest neighbor.

This method effectively forces the decision region of the minority class to become more general, rather than it being subsumed by the majority class samples around it. Han et al. [8] showed that most of the classification algorithms in the literature tried to learn the borderline of each class as exactly as possible in the training process. As the result, the samples on the borderline and those nearby were more easily misclassified than those far from the borderline. Based on the above analysis, they proposed two novel minority over-sampling methods, borderline-SMOTE1 and borderline-SMOTE2, which improved the performance of SMOTE by over-sampling minority class samples near the borderline. Chen et al. [9] proposed a novel hybrid resampling technique based on the differential evolution clustering hybrid re-sampling SVM algorithm (DEC-SVM), which utilized the mutation and crossover operators of differential evolution for over-sampling. Thus, by combining over-sampling and data cleaning techniques, only the useful samples remained, and computational efficiency was improved. Kubat et al. [10] showed samples of the majority class could have a detrimental effect on the learner's behavior, since noise or otherwise unreliable samples from the majority class could overwhelm the minority class. For this reason, an under-sampling method was proposed by removing noise and redundant majority class samples. The neighborhood cleaning rule (NCL) was presented by Laurikkala et al. [11], a technique that removed majority class samples based on Wilson's edited nearest neighbor rule [12]. The results suggested that NCL was a useful method for improving modeling of difficult imbalanced class problems.

In this paper, we focused on the problem of imbalanced class distribution in biomedical classification, including cancer classification from gene expression data. While SMOTE is a well-known general over-sampling method to address this problem, in some cases it cannot improve or can even reduce classification performance. Therefore, we developed a novel minority over-sampling method, named safe-SMOTE, in which only safe synthetic samples are generated so that any harmful effects of unsafe samples are suppressed.

In this paper, we will briefly introduce SMOTE and discuss its primary drawback, proposing a novel method, safe-SMOTE, based on this drawback (Section 2). We then present our findings and compare the results obtained using our novel method with those using the control (no over-sampling) and SMOTE methods (Section 3), before describing our conclusions (Section 4).

2. Methods

2.1 SMOTE

Chawla et al. developed a minority over-sampling technique called SMOTE (Synthetic Minority Oversampling TEchnique) [7] in which the minority class is over-sampled by creating synthetic samples rather than being over-sampled with replacement. SMOTE provided a new approach to over-sampling and introduced a bias towards the minority class. The results of their study showed that the SMOTE approach could improve the accuracy of classifiers for a minority class.

In SMOTE, the minority class is over-sampled by synthesizing new samples along the line between the minority samples and a random selection of their nearest neighbors. In a less application-specific manner, synthetic samples are generated by operating in "feature space" rather than "data space". Synthetic samples are generated by first computing the difference of the feature vector between each minority class sample and its selected nearest neighbor. Then, this difference is multiplied by a random number between 0 and l, and added to the feature vector of the minority sample. In this way, the synthetic minority sample is generated along the line segment between two specific features. Depending on the requirement of the over-sampling amount, the nearest neighbors are selected by chance.

This approach is effective in forcing the decision region of the minority class to become more general, as shown in Figure 1. Figure 1(A) presents a typical case of imbalanced data, where the samples from the majority class greatly outnumber those from the minority class. As a result, the majority class samples are well-classified, whereas many samples from the minority class are easily misclassified. Therefore, the problem of imbalanced datasets requires new and more adaptive methods, such as SMOTE. Figure 1(B) depicts how synthetic samples are generated by applying SMOTE in order to achieve a more balanced distribution, enabling the classifier to recognize all samples very well.

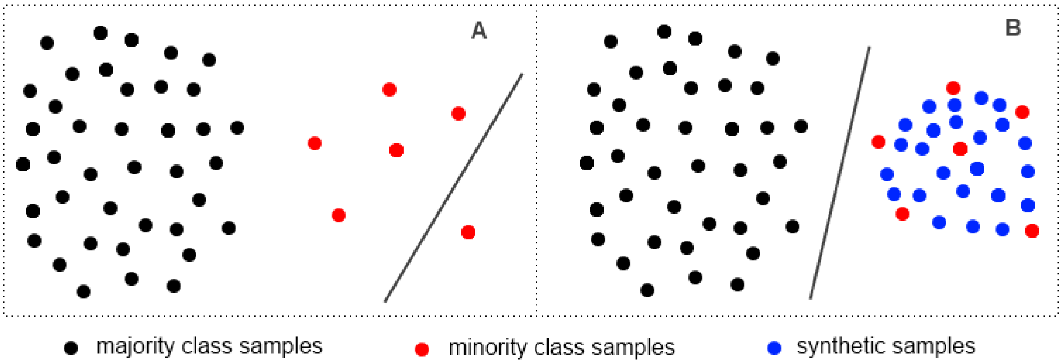

Figure 1. Advantages of SMOTE
Black, red, and blue dots indicate majority class, minority class, and synthetic minority class samples, respectively. The brown line depicts the discrimination hyperplane. (A) The original dataset with an erroneous classifier biased by the imbalanced dataset. (B) Synthesis of new minority class samples by applying SMOTE with a perfect classifier.

2.2 Main Drawback of SMOTE

To illustrate the above approach, Figure 2(A) shows how a given synthetic sample is generated using SMOTE. The blue sample *x* is a synthetic sample generated along the line that joins the minority class sample *s* and its randomly selected nearest neighbor *n*. Figure 2(B) presents a typical case of imbalanced data where the distribution of minority class samples is discrete and bordering the majority class samples. Figure 2(C) shows some synthetic minority samples generated using SMOTE; however, many of them are distributed inside and near the majority samples. Therefore, the classification accuracy could be reduced in comparison with the control method (i.e. no over-sampling). In Figure 2(C), the main drawback of SMOTE is apparent: many unsafe synthetic samples will not be paid any attention after they are generated. Therefore, in order to address this drawback and improve the classification accuracy of the SMOTE method, we focused on how to suppress the harmful effects of synthetic minority samples and only generate safe ones. This idea, illustrated in Figure 2(D), a novel method that we term safe-SMOTE, will be presented in more detail in the next section.

2.3 Safe-SMOTE

In order to overcome the drawback of SMOTE described previously, we focused on developing a way to suppress the harmful effect of synthetic minority samples and only generate safe ones. We note that *x* is generated by both *s* and *n*, where *n* is a randomly selected nearest neighbor whose position with respect to *s* is what determines *x*, as shown in Figure 2(A). If *s* is fixed, a change in *n* will lead to a change in *x*. Therefore, given a typical *s*, an unsafe *x* will be generated by an unsafe *n*, and a safe *x* will be generated by a safe *n*. Consequently, the question of how to suppress unsafe synthetic samples becomes replaced by that of how to suppress unsafe nearest neighbors.

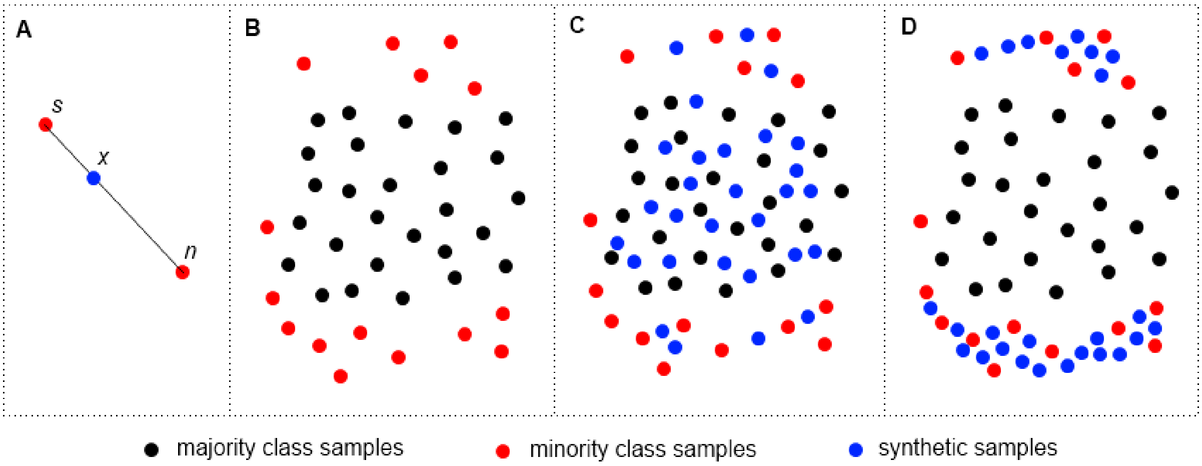

Figure 2. Main Drawback of SMOTE

Black, red, and blue samples are majority class, minority class, and synthetic minority class samples, respectively. (A) A representative synthetic minority class sample is generated along the line of two minority class samples by using SMOTE. (B) A typical case of imbalanced data. (C) Synthesis of new minority class samples by applying SMOTE. Many of these have harmful effects on the classifier. (D) Only safe synthetic samples are generated by using our novel method, safe-SMOTE.

Based on the above analysis, we propose a novel minority over-sampling method, safe-SMOTE. Safe-SMOTE adds one more module, the filter, into SMOTE to remove all unsafe nearest neighbors. Firstly, the center *o* of the hypersphere between *s* and *n* is determined. Then, the distances d_1 (between *o* and *s*) and d_2 (between *o* and each majority class sample *f*) are calculated. As shown in Figure 3(A), if there is at least one majority class sample inside the hypersphere (i.e. $d_1 > d_2$), we could say that *n* is an unsafe nearest neighbor for newly generated unsafe synthetic samples. Therefore, the filter will remove *n* from the nearest neighbors. On the other hand, as depicted in Figure 3(B), if all the majority class samples are outside the hypersphere (i.e. $d_1 < d_2$), *n'* will be called a safe nearest neighbor for generating the safe synthetic sample *x*.

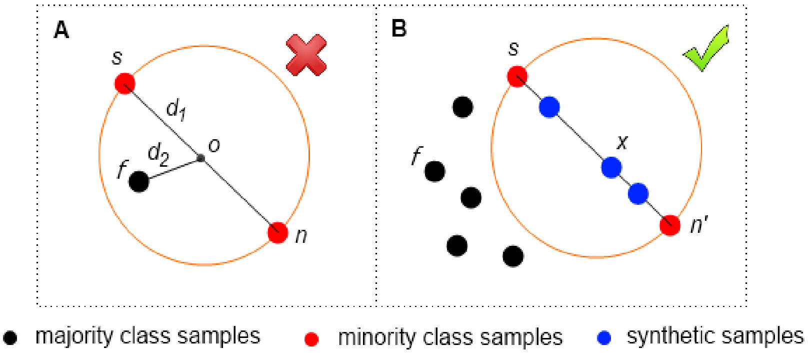

Figure 3. The filter for safe-SMOTE

Black, red, and blue samples are majority class, minority class, and synthetic minority class samples, respectively. (A) At least one majority class sample *f* stays inside the hypersphere between minority sample *s* and its nearest neighbor *n*; thus, *n* is the unsafe nearest neighbor and no synthetic samples are generated. (B) All majority class samples are outside the hypersphere; therefore, *n'* is a safe nearest neighbor. Consequently, this situation is appropriate to generate safe synthetic samples *x*.

The procedure of the safe-SMOTE method is explained briefly as follows:

```
Algorithm: safe-SMOTE(T,N,k)
Input: Number of minority class samples T; amount of safe-SMOTE N; number
       of nearest neighbors k;
Output: (N*T) synthetic minority class samples

I = (int)(N*T);
While (I > 0)
    Choose a minority sample ($s$). The choice is random but controlled
     to be as fair as possible.
    Calculate its k nearest neighbors.
    Apply filter to remove all dangerous nearest neighbors ($n$).
    Randomly select one of the safe nearest neighbors ($n'$).
    Synthesize new and safe $x$ along the line joining $s$ and $n'$.
    I := I - 1;
EndWhile;
Return a set of synthesized samples.
```

It should be noted that the safety for the above procedure has been considered only in the feature space. This means that if we use a kernel method like SVM, the safety is not guaranteed in the kernel space. It would be interesting to develop a “kernel-safe-SMOTE” method; however, this would involve losing the freedom of choosing difference classifiers.

2.4 Evaluation measures

In the medical science, bioinformatics, and machine learning communities [1][13][14][15], sensitivity (SE) and the specificity (SP) are two metrics used to evaluate the performance of classifiers. SE measures the proportion of actual positives that are correctly identified as such, while SP can be defined as the proportion of negatives that are correctly identified. Kubat et al. [10] proposed the geometric mean (G-mean) metric as defined below.

$$\text{G - mean} = \sqrt{SE \times SP}$$

It is common practice to apply this metric to evaluate classifiers used in imbalanced class distributions [1][10][15][16][17], and so we also decided to use it to measure the performances of the classifiers in our research. Additionally, we also calculated the F-measure as another important metric.

2.5 Classifier

Support Vector Machine (SVM) is a supervised learning machine widely used to build a classifier that discriminates classes for binary class classification [18]. SVM is based on simple ideas originating from statistical learning theory [19] and has high generalization capability, optimizes global classification solutions, and can be successfully applied in bioinformatics.

In this study, an SVM implementation in the kernlab package [20] was used, available at the Comprehensive R Archive Network (CRAN). This is an extensible package for kernel-based machine learning methods in R and includes various kernels such as the Linear kernel, Polynomial kernel, and Radial Basis kernel (Gaussian kernel), the latter of which was employed in this study. Using heuristics, kernlab automatically optimizes the value of the sigma parameter for the Radial Basis kernel to achieve better classification performance in most practical situations. In addition, all other hyper-parameters, e.g. cost and class weights, are set to default values.

3. Experiments and Discussions

3.1 Datasets

In this study, the cancer classifications from gene expression data selected were colon-cancer, as presented by Alon et al. [21], and acute leukemia, as described by Golub et al. [22]. The colon-cancer dataset consisted of 62 colon tissue samples (22 normal and 40 tumor) with 2000 features, while the leukemia dataset consisted of 72 patients (25 acute myeloid leukemia patients (AML) and 47 acute lymphocytic leukemia patients (ALL)) with 7,129 features. The former dataset was considered positive, while the latter was negative. The positive-to-negative class imbalance ratios of the two datasets were 0.35:0.65 and 0.34:0.66, respectively.

Table 1. Description of the datasets

Name	Examples	Attributes	Imbalance ratio
*colon-cancer	62	2000	0.35 : 0.65
*leukemia	72	7129	0.34 : 0.66
ionosphere	351	34	0.36 : 0.64
pima	768	8	0.35 : 0.65
breast-w	683	10	0.35 : 0.65
blood	748	4	0.23 : 0.77
satimage	6435	36	0.097 : 0.903
yeast	1484	8	0.034 : 0.966

In order to demonstrate the applicability of our method, we also performed experiments using six real-world imbalanced benchmark datasets obtained from UCI [23]: Radar data (ionosphere), Pima Indians Diabetes (pima), Breast Cancer Wisconsin (breast-w), Blood Transfusion Service Center (blood), Landsat Satellite (satimage), and Yeast (yeast), each with a different class imbalance ratio, as shown in Table 1. For highly imbalanced problems, the classes “damp grey soil” and “ME2” of the satimage and yeast datasets, respectively, were converted into the minority class and the remaining classes of each dataset became the majority class. Except for ionosphere and satimage, these datasets all contained biomedical data.

3.2 Classification imbalance learning results

The experiments were executed to compare three methods: the control method (no over-sampling), SMOTE, and safe-SMOTE. A SVM was used as the classifier. The classification performances of the methods were all estimated based on the 10-fold cross-validation strategy. For each test, nine-tenths of the complete dataset were used as a training set. Then, for the cases of SMOTE and safe-SMOTE, minority samples in the training set were over-sampled. After training by an SVM model using the (possibly over-sampled) training set, the model was tested against the remaining one-tenth of the dataset (i.e. test set). This process was repeated 10-fold for all datasets and methods with different combinations of training and test sets. The values for the performance criteria—SE, SP, G-mean, and F-measure—were calculated by averaging 20 independent runs of 10-fold cross-validation: they are summarized in Tables 2 and 3. Furthermore, two-sample t-tests with equal variance were conducted to assess whether the averages of the G-mean and F-measure by different methods were significantly different.

Experimental results from the two gene expression datasets for cancer classification (leukemia and colon-cancer) showed that our method (i.e. safe-SMOTE) achieved a better G-mean than both the control method and SMOTE. For example, in the colon-cancer dataset, although the sensitivity and specificity increased for the control method (81.59% and 89.50%) by SMOTE (81.36% and 88.63%), it was also increased (to 81.82% and 90.50%) by safe-SMOTE. Furthermore, the G-mean of the control (85.45%) was reduced to 84.91% by SMOTE, but increased by safe-SMOTE (86.04%). In the leukemia dataset, the sensitivity and G-mean of the control method (54.80% and 74.00%) were improved by SMOTE (78.60% and 88.64%); but, safe-SMOTE achieved even higher performance for these two criteria (80.80% and 89.82%). However, a different case was observed for the specificity in the leukemia dataset, with the specificity of the control method (100.00%) being unchanged when using SMOTE (100.00%), but decreasing slightly by 0.11% for safe-SMOTE (99.89%).

Table 2. Comparison of Sensitivity (SE) and Specificity (SP) (%)

	SE			**SP**		
	no over-sampling	**SMOTE**	**safe-SMOTE**	**no over-sampling**	**SMOTE**	**safe-SMOTE**
*colon-cancer	81.59	81.36	**81.82**	89.50	88.63	**90.50**
*leukemia	54.80	78.60	**80.80**	**100.00**	**100.00**	99.89
ionosphere	89.96	**94.52**	93.73	**97.00**	93.87	96.07
pima	55.11	**82.26**	81.14	**87.45**	67.21	69.69
breast-w	98.44	98.71	**99.40**	94.13	95.12	**95.51**
blood	30.65	**74.04**	69.27	**94.21**	60.35	66.96
satimage	51.26	85.30	**87.60**	**97.99**	92.62	91.94
yeast	3.73	48.82	**50.39**	**100.00**	97.01	97.03

Table 3. Comparison of G-mean and F-measure (%)

	G-mean			**F-measure**		
	no over-sampling	**SMOTE**	**safe-SMOTE**	**no over-sampling**	**SMOTE**	**safe-SMOTE**
*colon-cancer	85.45	84.91	**86.04**	81.32	80.56	**82.22**
*leukemia	74.00	88.64	**89.82**	70.75	87.98	**89.25**
ionosphere	93.41	94.19	**94.89**	92.12	92.01	**93.38**
pima	69.42	74.35	**75.19**	61.74	67.58	**68.27**
breast-w	96.26	96.90	**97.44**	93.93	94.92	**95.61**
blood	53.71	66.84	**68.10**	41.07	49.19	**50.37**
satimage	70.88	88.88	**89.74**	60.35	**67.22**	66.76
yeast	17.93	68.81	**69.91**	7.11	41.93	**43.13**

Assessment by t-tests suggested that in the colon-cancer dataset, the control method significantly outperformed SMOTE (p = 3.44E-2), but safe-SMOTE achieved a significantly higher G-mean than both the control method and SMOTE could (p = 2.25E-2 and p = 9.86E-4, respectively). Furthermore, in the leukemia dataset, SMOTE significantly surpassed the control method (p = 2.2E-16) for G-mean, but safe-SMOTE significantly outperformed them both (p = 2.2E-16 and p = 2.95E-2, respectively).

We also performed experiments using the six imbalanced benchmark datasets from UCI to demonstrate the general applicability of our method. The experimental results also show our method as having achieved higher G-mean values than both the control and SMOTE methods in all six datasets. Furthermore, two-sample t-tests showed that our method remarkably outperformed both the control and SMOTE methods ($p < 0.05$; see Table 4).

Table 4. The assessment by two-sample t-test with equal variance

Dataset	Compared methods	P-value for G-mean	P-value for F-measure
colon-cancer	no over-sampling vs SMOTE	3.44E-02	3.40E-02
	safe-SMOTE vs no over-sampling	2.25E-02	1.90E-02
	safe-SMOTE vs SMOTE	9.86E-04	1.10E-03
leukemia	SMOTE vs no over-sampling	2.20E-16	2.20E-16
	safe-SMOTE vs no over-sampling	2.20E-16	2.20E-16
	safe-SMOTE vs SMOTE	2.95E-02	3.40E-02
ionosphere	SMOTE vs no over-sampling	6.24E-05	6.70E-01
	safe-SMOTE vs no over-sampling	3.22E-09	5.99E-06
	safe-SMOTE vs SMOTE	5.69E-05	6.80E-08
pima	SMOTE vs no over-sampling	2.20E-16	2.20E-16
	safe-SMOTE vs no over-sampling	2.20E-16	2.20E-16
	safe-SMOTE vs SMOTE	1.76E-05	6.09E-04
breast-w	SMOTE vs no over-sampling	1.81E-10	3.83E-12
	safe-SMOTE vs no over-sampling	2.20E-16	2.20E-16
	safe-SMOTE vs SMOTE	8.79E-10	9.84E-09
blood	SMOTE vs no over-sampling	2.20E-16	4.67E-16
	safe-SMOTE vs no over-sampling	2.20E-16	2.20E-16
	safe-SMOTE vs SMOTE	6.38E-06	8.60E-05
satimage	SMOTE vs no over-sampling	2.20E-16	2.20E-16
	safe-SMOTE vs no over-sampling	2.20E-16	2.20E-16
	safe-SMOTE vs SMOTE	2.80E-12	4.23E-05
yeast	SMOTE vs no over-sampling	2.20E-16	2.20E-16
	safe-SMOTE vs no over-sampling	2.20E-16	2.20E-16
	safe-SMOTE vs SMOTE	1.30E-02	1.40E-02

4. Conclusions

In this paper, we have addressed a common problem in efforts to classify cancers from gene expression data, known as the imbalanced class distribution problem. To this end, we proposed a novel minority over-sampling method called safe-SMOTE, which is an improved version of the SMOTE method in which only safe synthetic samples are generated, thereby suppressing the harmful effect of unsafe ones.

Experimental results obtained from six imbalanced benchmark datasets from the UCI Machine Learning Repository and two gene expression datasets for cancer classification showed that our method achieved better G-mean and sensitivity than both the control and SMOTE methods ($p < 0.05$). These results suggest that our method can outperform SMOTE in various biomedical classification problems, including cancer classification.

Although safe-SMOTE achieved improvements and high performances in cancer classification, several further directions remain to be considered. These include combining our novel method with feature selection methods, applying other novel under-sampling methods for cancer classification, and extracting new and appropriate sets of features from gene expression data. Addressing these challenges will be a key aim of our work in the future.

In conclusion, a number of improvements to SMOTE have recently been reported. These have included a hybrid preprocessing approach based on SMOTE and Rough Set Theory (SMOTE-RSB) [24], a novel probability density function estimation-based over-sampling SMOTE [25], and the combination of a heuristic-based unsupervised feature selection technique and SMOTE [26]. The method proposed in our study, safe-SMOTE, differs markedly from such methods, and it may be interesting and worthwhile to consider its use in combination with these existing approaches.

References

[1] Batuwita, R.; Palade, V. microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics* **2009**, *25(8)*, 989-95.

[2] Radivojac, P. et al. Classification and knowledge discovery in protein databases. *Journal of Biomedical Informatics* **2004**, *37(4)*, 224-239.

[3] Yu, C.Y.; Chou, L.C.; Chang, D. T. H. Predicting protein-protein interactions in unbalanced data using the primary structure of proteins. *BMC Bioinformatics* **2010**, *11*, 167.

[4] Joshi, M. V.; Kumar, V.; Agarwal, R. C. Evaluating boosting algorithms to classify rare classes: comparison and improvements, Proceedings 2001 IEEE International Conference on Data Mining, 2001, pp 257-264.

[5] Lin, Y.; Lee, Y.; Wahba, G. Support Vector Machines for Classification in Nonstandard Situations by in Nonstandard Situations. *Machine Learning* **2002**, *46(1-3)*, 191-202.

[6] Wu, G.; Chang, E. Y. Class-Boundary Alignment for Imbalanced Dataset Learning, In Proc. of the ICML'03 Workshop on Learning from Imbalanced Data Sets, 2003, pp 49-56.

[7] Chawla, N. V.; Bowyer, K. W.; Hall, L. O. SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* **2002**, *16*, 321-357.

[8] Han, H.; Wang, W.-yuan.; Mao, B.-huan. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning, In *Advances in Intelligent Computing, Lecture Notes in Computer Science Volume 3644*; Springer: Germany, 2005; pp 878-887.

[9] Chen, L.; Cai, Z.; Chen, L. A Novel Differential Evolution-Clustering Hybrid Resampling Algorithm on Imbalanced Datasets, 2010 Third International Conference on Knowledge Discovery and Data Mining, 2010, pp 81-85.

[10] Kubat, M.; Matwin, S. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection, Proceedings of the Fourteenth International Conference on Machine Learning, 1997, pp 179-186.

[11] Laurikkala, J. Improving Identification of Difficult Small Classes by Balancing Class Distribution, In *Artificial Intelligence in Medicine*, *Lecture Notes in Computer Science Volume 2101*; Springer: Germany, 2001; pp 63-66.

[12] Wilson, D. R.; Martinez, T. R. Reduction Techniques for Instance-Based Learning Algorithms. *Machine Learning* **2000**, *38(3)*, 257-286.

[13] Akbani, R.; Kwek, S.; Japkowicz, N. Applying Support Vector Machines to Imbalanced Datasets, In *Machine Learning: ECML 2004, Lecture Notes in Computer Science Volume 3201*; Springer: Germany, 2004; pp 39-50.

[14] Xiao, J.; Tang, X.; Li, Y.; Fang, Z.; Ma, D.; He, Y.; Li, M. Identification of microRNA precursors based on random forest with network-level representation method of stem-loop structure. *BMC Bioinformatics* **2011**, *12(1)*, 165.

[15] Anand, A.; Pugalenthi, G.; Fogel, G. B.; Suganthan, P. N. An approach for classification of highly imbalanced data using weighting and undersampling. *Amino Acids* **2010**, *39(5)*, 1385–91.

[16] Han, K. Effective sample selection for classification of pre-miRNAs. *Genetics and molecular research* **2011**, *10(1)*, 506–18.

[17] Xuan, P. et al. PlantMiRNAPred: efficient classification of real and pseudo plant pre-miRNAs. *Bioinformatics* **2011**, 27(10), 1368–76.

[18] Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery* **1998**, *2*, 121–167.

[19] Vapnik, V. N. An overview of statistical learning theory. *IEEE Transactions on Neural networks* **1999**, *10(5)*, 988–99.

[20] Karatzoglou, A.; Smola, A.; Hornik, K. kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software* **2004**, *11(9)*.

[21] Alon, U. et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays., Proceedings of the National Academy of Sciences of the United States of America, 1999, *96(12)*, pp 6745–6750.

[22] Golub, T. R. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* **1999**, *286(5439)*, 531–537.

[23] Frank, A.; Asuncion, A. UCI Machine Learning Repository, [http://archive.ics.uci.edu/ml], **2010**, Irvine, CA: University of California, School of Information and Computer Science.

[24] Ramentol, E.; Caballero, Y.; Bello, R. 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* **2012**, *33(2)*, 245–265.

[25] Gao, M.; Hong, X.; Chen, S.; Harris, C. J. Probability Density Function Estimation Based Over-Sampling for Imbalanced Two-Class Problems, International Joint Conference on Neural Networks, 2012, pp 1–8.

[26] Kerdprasop, N.; Kerdprasop, K. On the Generation of Accurate Predictive Model from Highly Imbalanced Data with Heuristics and Replication Techniques. *International Journal of Bio-Science and Bio-Technology* **2012**, *4(1)*, 49–64.