BSMOTE with LDA for High-Dimensional and Class-Imbalanced Ovarian Cancer Data

Siyuan He

College of Information Science and Technology, Jinan University, Guangzhou 510632, China;

hesiyuanlk@163.com

Abstract

In order to solve class-imbalance problem of high-dimensional ovarian cancer, a feasible method is proposed. It is difficult to classify the patients and healthy controls when an imbalance data set with not enough samples of patients or healthy controls is used. However, expanding samples is difficult to implement because of high cost of finding satisfactory participants. Synthetic Minority Over-Sampling Technique (SMOTE) is a practical over-sampling method to synthesize new samples of minority class. In this paper, an improved SMOTE which named Borderline Synthetic Minority Over-Sampling Technique (BSMOTE) is combined with Linear Discriminant Analysis (LDA) to improve the classification accuracy for ovarian cancer data. Compared with Principal Component Analysis (PCA), LDA has better performance in several criteria.

Keywords

Class-Imbalance, Dimension Reduction, Borderline Synthetic Minority Over-Sampling Technique, Principal Component Analysis, Linear Discriminant Analysis.

1. Introduction

Proteomics is a science that takes the proteome as a research object and studies the protein composition and its variation law of cells, tissues or organisms [1]. The term proteomics derived from a combination of two terms, protein and genome, means "A proteonle is the entire PROTein complement expressed by a genOME," which includes all proteins expressed by one cell or even one organism[2]. Proteomics essentially refers to the study of the characteristics of large scale proteins, including the expression level of proteins, post-translational modifications, protein-protein interactions and the like, thereby obtaining information on disease states, cell metabolism, etc., at the protein level.

In this paper, proteomics expression data is our main research object which includes the analysis of large scale proteins. It helps identify main proteins in a particular sample, and those proteins differentially expressed in related samples, such as diseased and healthy tissues. If a protein is found only in a diseased sample then it can be a useful drug target or diagnostic marker[3]. Proteins with same or similar expression profiles may also be functionally related so protein data analysis can effectively diagnose some diseases such as ovarian cancer[4,5].

However, in machine learning, the situation of class-imbalance has been reported hindering the performance of some standard classifiers such as Support Vector Machines (SVM) [6]. The classifiers which do not consider imbalanced class will attach importance to majority class and ignore the minority class[7]. For example, a data set has 90 samples of cancer patient and only one 10 samples of healthy control. To minimize the error rate, learning algorithm may classify all examples as the majority class and achieve a low error rate of 10%. In this way, all minority class examples will be wrongly classified. Especially in medical science field, data set of imbalanced class appear more frequently such as cancer data set. In some cases we cannot obtain balanced data set and the number of different classes of samples may vary widely. Learning from imbalanced data sets usually produces biased classifiers and results in higher predictive accuracy of majority class but poorer predictive accuracy of the minority class.

To solve this problem, SMOTE was proposed for over-sampling. However, SMOTE dose not considers the distribution of samples and behaves worse. With BSMOTE, only the part of minority class samples which have closer Euclidean distance with majority class samples (means minorities samples of danger zone) will be over-sampled.

In addition, dimension reduction is necessary for these high-dimensional protein profile. Dimension reduction is another challenge because the method of over-sampling will change the covariance of minority class sample. Covariance is an important factor of dimension reduction method like PCA and LDA. PCA is an unsupervised method when LDA needs tag for each class. Our target is to choose a feasible dimension reduction method for over-sampled data and compare their performance in several aspects.

As we know, small data usually results in under-fitting. The difficulty of classification is sorted as follows, Big Data + Balanced Data < Big Data + Imbalanced Data < Small Data + Balanced Data < Small Data + Imbalanced Data. The purpose of this article is to select feasible classification method for high-dimensional and imbalanced data such as ovarian cancer protein profile, so we paid our attention on Small Data + Imbalanced Data. Our contributions include,

combine the BSMOTE with dimension reduction method to improve the performance of classification for ovarian cancer data;

compare unsupervised method PCA with supervised method LDA to show the advantage of LDA for high-dimensional ovarian data set after over-sampling;

conduct experiments on data from different platforms to enhance the persuasion of results.

The main idea of our method can be vividly described in Figure 1, different classes of raw data can be regarded as two stacks of books with different quantities. Firstly, We balance both and then extract the contents of each book to pages. The classifier only needs to learn from these pages instead of books that full of redundant information.

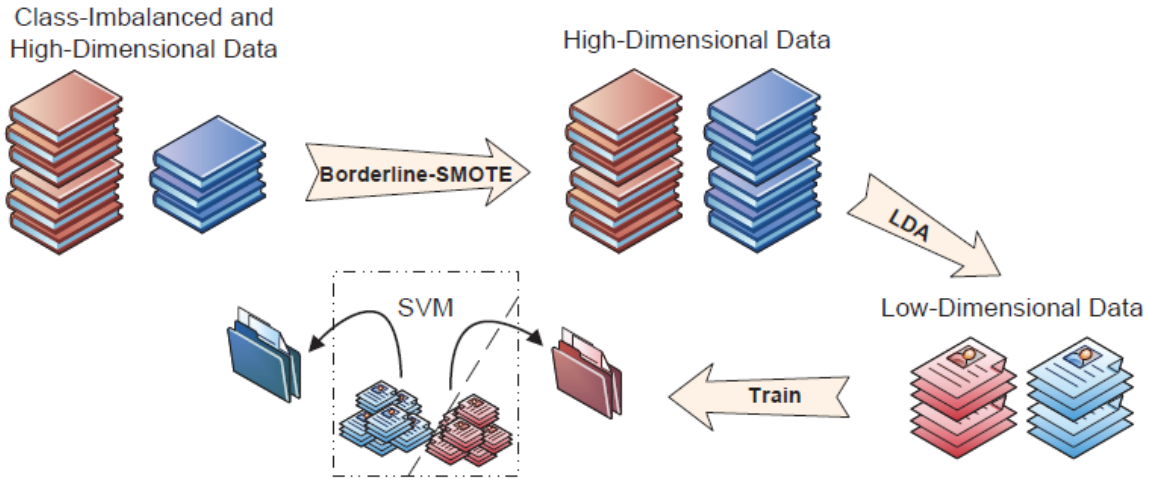


Fig. 1 Main idea of our method

The rest of this article is organized as follows. Section 2 introduces the related work. Section 3 firstly describes the details of data sets ,then mainly introduces the methods of BSMOTE, PCA and LDA, respectively. In section 4 we propose our method and conduct experiment. The empirical results are showed in section 5 and our conclusion is presented in Section 6.

2. Related work

Biological interactions networks are very complicated and contain a lot of variables [8]. In the biomedical field, the expression profiles of genes, miRNAs or proteins are commonly measured by high-dimensional tools such as microarray. These high-dimensional data sets are obtained from high-throughput technologies which can measure numerous genes for each sample. However, the number

of variables greatly exceeds the number of samples [9]. In this case, the classifier is often difficult to get effective training.

Class-imbalanced data is another common problem in the biomedical field [10]. Some works explored the effect of imbalanced data on existing methods. Xie and Qiu demonstrated that the imbalanced data sets have a negative effect on the performance of LDA theoretically. The experimental results also show that over-sampling methods are more effective than under-sampling methods in improving the performance of LDA [11]. Furthermore, a new over-sampling method SMOTE can be used for achieving better true positive (TP) rate and F-value than random over-sampling or SMOTE methods [12]. Considering the impact of imbalance data set on SVM and linear proximal support vector machine (LPSVM) , Zhuang *et al* proposed Weighted LPSVM (WLPSVM) in order to deal with class-imbalance problem[13]. However, WLPSVM focuses on the sample numbers of different classes instead of the sample distribution.

Considering the distribution of samples, BSMOTE over-samples the samples of danger zone, balances the quantity and distribution which more conducive to classification [14]. Reference [15] studied the advantages of utilizing deep learning techniques for feature analysis and transfer learning. Their framework alleviates the issue of imbalanced data. Deep learning for imbalanced data is a state-of-the-art method depending on big data [16]. Reference [17] proposed Deep Over-sampling (DOS) to extend the synthetic over-sampling method which exploiting deep feature space by convolutional neural networks (CNNs). Reference [18] researched the impact of class imbalance on CNNs and compare CNNs with frequently used methods for class-imbalanced data. In our study, we focus on small data in the field of biomedicine and synthesis samples by over-sampling algorithm. Considering the covariance would be changed after over-sampling, different dimension reduction algorithms are used in this paper.

3. Material and Methods

3.1 Data Set

The protein profile data set used in our study is downloaded from National Center for Biotechnology Information (NCBI) and the data set number is GSE79517. It was formed via Nucleic Acid Programmable Protein Array (NAPPA) Serous Ovarian Discovery Array and consisted of five data sets from different platforms. The first data we used from the platform GPL21630 consists of 85 sample (including 28 samples of healthy control, 29 samples of benign ovarian disease and 28 samples of serous disease). The second data from the platform GPL21624 consists of 43 samples (including 28 serous disease and 15 healthy control).

The dimension in feature space of the raw data is very high. Each sample in the first data $Data_1$ and $Data_2$ has 1469 features and 1760 features respectively. Our research is aimed at high-dimensional and class-imbalanced cancer data like $Data_1$ and $Data_2$.

3.2 Over-sampling Method BSMOTE

Since the positive and negative samples in the data are imbalanced, a new over-sampling method named BSMOTE is used to process the data. Standard SMOTE is based on a random over-sampling algorithm to improve the performance. The problem of random over-sampling which takes a simple copy of the sample to increase the number of samples is so easy to produce the over-fitting problem and the information learned from model is too special.

The basic idea of BSMOTE algorithm is to analyse the minority class and add new samples to a data set based on the few samples near the borderline. It should be noted that the examples far from the borderline between different classes contribute little to classification and the examples near the borderline are easily misclassified, so only the samples of minority class which near the borderline will be considered. Furthermore, BSMOTE can be subdivided into BSMOTE1 and BSMOTE2. The difference between them is that BSMOTE2 not only syntheses each minority samples in DANGER zone and its positive nearest neighbors, but also does that from its nearest negative neighbors [12].

In this study, we mainly use the BSMOTE2 for experiment, and the BSMOTE we use in the remaining part refers to BSMOTE2.

Suppose the set of majority class is $N=\{n_i, i=1,2,...,nnum\}$ and the set of minority class is $P=\{p_i, i=1,2,..., pnum\}$, the number of minority and majority samples are $pnum$ and $nnum$. The procedure of BSMOTE is as follows.

Firstly, k nearest neighbors from training set T (contains minority and majority classes) are calculated for every sample p_i in P. And the number of majority examples among these k nearest neighbors is recorded as $k'(0\leqslant k'\leqslant k)$.

Secondly, for one of the samples p_i in P, if $k'=k$,it means that all of its nearest neighbors come from N so this sample will be regarded as noise and discarded. If $k/2\leqslant k'<k$, the number of its nearest neighbors from P is less than those from N. Samples like this will be considered to be easily misclassified and put into the set $DANGER$. If $0\leqslant k'<k/2$, p_i contributes little to classification and needs not to take part in the follows steps.

The set $=\{d_i, i=1,2,...,dnum\}(0\leqslant dnum\leqslant pnum)$ saves the borderline samples from set P, in other words $d_i\subseteq P$. For each d_i we calculate its k nearest neighbors from P and N.

In this step, $s\times dnum$ examples are generated from the $DANGER$, where s refers to the sampling rate $(1\leqslant s\leqslant k)$. For each d_i, we randomly select s nearest neighbors from its k nearest neighbors in P and N. Then the distance between d_i and its nearest neighbors are saved as $dif_j(j=1,2,...,s)$. The last step is to multiply dif_j by a random number between 0 and 0.5 and the new synthetic samples are generated between d_i and its nearest neighbors,

$$synthetic_j = d_i + rand(0,0.5)\times dif_j. \tag{1}$$

3.3 Feature Extraction by using PCA

High-dimensional and curse of dimensionality are common phenomenons in expression proteomics data and over-fitting is direct manifestation of these problems [19], so it is the reason why we put principal component analysis on application [20]. The principle of PCA is to project a high-dimensional vector x into a low-dimensional vector space to obtain a low-dimensional vector y. Vector y contains orthogonal variables called principal components. Through the low-dimensional representation of the vector and eigenvector matrix, we can basically reconstruct the corresponding original high-dimensional vector.

Assuming that there are n genetic training samples with m dimension, m-dimension vector x_i contains the values of each sample and the set of training samples can be defined as $X=\{x_1, x_2,..., x_n\}$. The average vector of the set is

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i, \tag{2}$$

and the covariance matrix of the set is

$$\hat{\Sigma}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^{\mathrm{T}}, \tag{2}$$

then calculate the eigenvector u_i of the covariance matrix and the corresponding eigenvalue. The eigenvalues consisted of the covariance matrix are sorted in descending order $\{\lambda_1,\lambda_2,...,\lambda_m\}$. Suppose that the corresponding eigenvectors which greater than or equal to λ_d are regarded as the principal

component, we can obtain the subset $\{\lambda_1, \lambda_2, ..., \lambda_d\}$. The corresponding transformation matrix of the principal component is

$$U = (u_1, u_2, ..., u_d), \tag{4}$$

so that each sample of the genetic data can be projected into a feature subspace, and the shape of U is $n \times d$. In this way, for any sample we have

$$y = U^{\mathrm{T}}(x - \overline{x}). \tag{5}$$

y is a low dimension vector which we obtain through the PCA algorithm. For our experiment data set, some of the gene expression information may be lost but dose not affect the overall quality.

3.4 Supervisory Dimensionality Reduction LDA

Linear Discriminant Analysis is a commonly used technique for dimension reduction. For any particular data, LDA maximizes the differences of variance between within-class and between-class to ensure maximal separability [21]. Suppose that our data set is $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$. Any sample x_i is an m-dimensional vector and $y_i \in \{0,1\}$, where m is the number of dimension of samples. We define $X_j (j = 0,1)$ as the set of class j and $\mu_j (j = 0,1)$ as the mean vector of class j, and $\hat{\Sigma}_j (j = 0,1)$ as the covariance matrix of class j.

Assuming that our projection plane is an d-dimensional vector w, then for any sample x_i it has a projection $w^{\mathrm{T}} x_i$ on vector w. For our two classes center points μ_0, μ_1, $w^{\mathrm{T}}\mu_0$ and $w^{\mathrm{T}}\mu_1$ are the projections on the w. Since the LDA needs to make the distance between the category centers μ_0, μ_1 as large as possible. In order to achieve the goal we maximize $\| w^{\mathrm{T}}\mu_0 - w^{\mathrm{T}}\mu_1 \|_2^2$. We want the projection points of same class data to be as close as possible. That is, the covariances $w^{\mathrm{T}}\hat{\Sigma}_0 w$ and $w^{\mathrm{T}}\hat{\Sigma}_1 w$ should be as small as possible. Thus we minimize $w^{\mathrm{T}}\hat{\Sigma}_0 w + w^{\mathrm{T}}\hat{\Sigma}_1 w$. In general, our optimization target is

$$\text{maximize } J(w) = \frac{\| w^{\mathrm{T}}\mu_0 - w^{\mathrm{T}}\mu_1 \|_2^2}{w^{\mathrm{T}}\hat{\Sigma}_0 w + w^{\mathrm{T}}\hat{\Sigma}_1 w} = \frac{w^{\mathrm{T}}(\mu_0 - \mu_1)(\mu_0 - \mu_1)^{\mathrm{T}} w}{w^{\mathrm{T}}(\hat{\Sigma}_0 + \hat{\Sigma}_1) w}. \tag{6}$$

Within-class scatter matrix S_w is defined as

$$S_w = \hat{\Sigma}_0 + \hat{\Sigma}_1 = \sum_{x \in X_0} (x - \mu_0)(x - \mu_0)^{\mathrm{T}} + \sum_{x \in X_1} (x - \mu_1)(x - \mu_1)^{\mathrm{T}} \tag{7}$$

and the between-class scatter matrix S_b is defined as

$$S_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^{\mathrm{T}}. \tag{8}$$

Therefore our optimization target can be reconstructed as

$$\text{maximize } J(w) = \frac{w^{\mathrm{T}} S_b w}{w^{\mathrm{T}} S_w w}. \tag{9}$$

According to the generalized rayleigh quotient [22], the maximum of $J(w)$ is the largest eigenvalue of the matrix $S_w^{-1} S_b$ and the w is the eigenvector corresponding to the largest eigenvalue. According to the definition of eigenvector,

$$(S_w^{-1} S_b) w = \lambda w. \tag{10}$$

For two-class problem, the direction of $S_b w$ is $\mu_0 - \mu_1$. Replace $S_b w$ with $\lambda(\mu_0 - \mu_1)$ in Equation 10, so we have

$$w = S_w^{-1}(\mu_0 - \mu_1). \tag{11}$$

To obtained the best projection direction w, we mainly need to calculate the mean and variance of samples from two different classes, respectively.

3.5 SVM Model

Proteomics expression data is an importance component for clinical decision support and plays a key role in prediction of clinical outcomes of involuted diseases such as cancer. To maximize the benefits of this technology, researchers are continuously looking for algorithms with the most accurate for the creation of proteomics expression profiles. Prior research suggests that SVM achieves better classification performance than back propagation neural networks, K-nearest neighbors, probabilistic neural networks, decision trees and weighted voting methods. SVM is regarded as the most promising techniques for proteomics expression data classification [23].

In this article we use SVM for classification and the solution to SVM amounts to the solution to Quadratic Programming (QP) optimization problem. Sequential minimal optimization (SMO) which was presented by John C. Platt has been widely used for SVM training [24]. The training problem is broke into a series of smallest possible sub-problems by SMO and then solved analytically. SVM comes to be the most popular classification technology benefited from the efficient calculating of SMO.

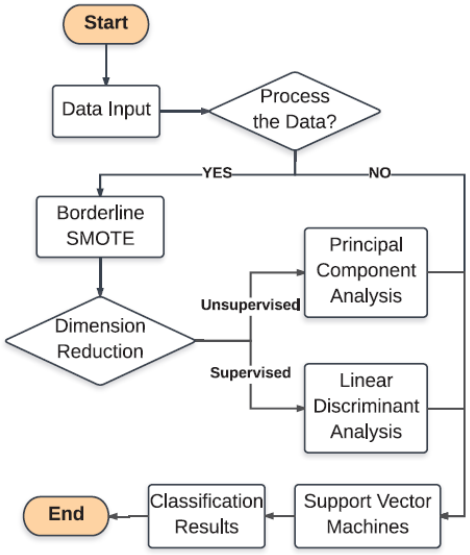


Fig. 2 Frame of experiment

4. Experimental methodology

The main purpose of our experiment is to compare the performance of our BSMOTE-LDA with other exciting method. A common method of calculating the area under the ROC curve (AUC) is used to make this comparison purpose [25]. The reason why we used ROC is not only ROC combines sensitivity and specificity but also accurately reflects the relationship between them [26]. AUC is a common method for classifiers comparing [27]. In this study we use ipython 5.3.0 (an enhanced interactive python) to build our classifier and process the data. In addition, we conduct cross-validation based on the data set. Then process data by training set and classify with testing set. The steps of the experiment are as follows.

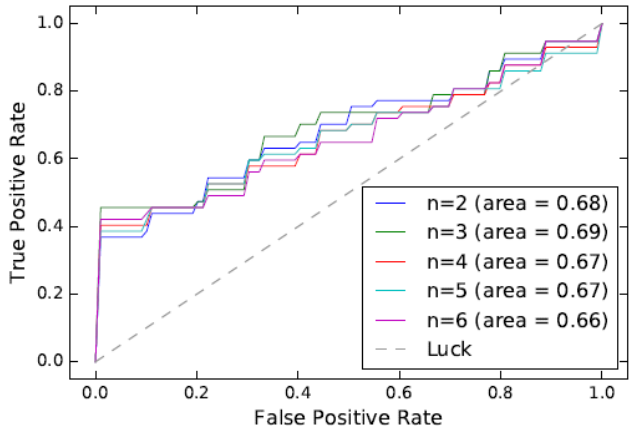
*Step*1: We input the training set, testing set and directly get the results of Only-SVM though standard SVM. Then we use the Borderline Synthetic Minority Oversampling Technique to generate the samples of minority class in training set. In this part, we set the sampling rate to 100% and the number of its nearest neighbors is 3. So we can obtain data which hold the same proportion between minority and majority classes. Then we save the data after over-sampling for the second step.

*Step*2: The dimension of ovarian cancer data is reduced by different methods in this part. We set the number of components to 1 for both PCA and LDA so we obtain two different data after dimension reduction. And then we employ SVM to establish the detection model and train the SVM model with linear kernel for these low-dimensional data. In this way we get another two results of BSMOTE-PCA1 and BSMOTE-LDA1.

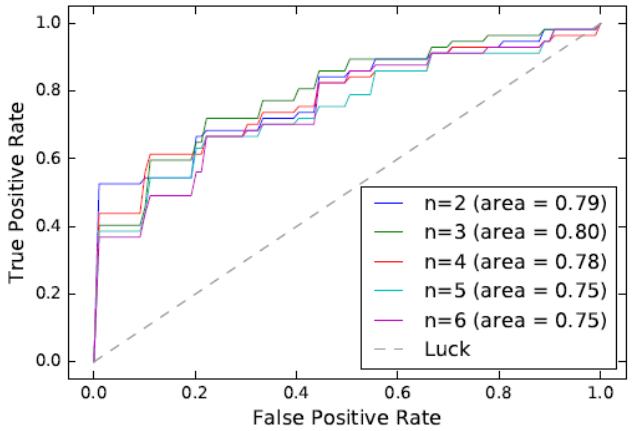
*Step*3: In order to compare with PCA in different principal components, we set the number of component to be 50 (which has the best performance of PCA in our experiment). After that, we repeat the process of the second steps and obtain the result of BSMOTE-PCA50.

Figure 2 shows the frame of our experiment. According to different choices in experiment, four results of different methods are obtained. Since we have two different data sets ($Data_1$ and $Data_2$), we conduct the same operation for these data.

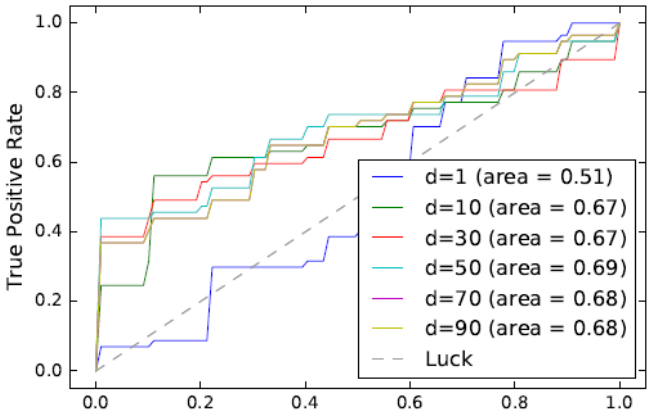
Figure 3 shows the effects of methods display in different parameters, illustrating the reason we choice these parameters for our experiment. Further more, to explore the efficiency of BSMOTE-LDA1, we conduct experiment to compare BSMOTE-LDA1 with Only-SVM, BSMOTE-PCA50 and BSMOTE-PCA1 in different classifiers. Besides SVM, Random Forest (RF) [28], Decision Tree (DT) [29] and k-Nearest Neighbor (KNN) [30] are applied in our experiment to show the performance of BSMOTE-LDA1.



(a) Different NN for BSMOTE-PCA



(b) Different NN for BSMOTE-LDA



(c) Different dimensions for BSMOTE-PCA

Fig. 3: Different number of Nearest Neighbor (NN) and dimensions can cause the difference of performance. From these average results of 3-fold cross-validation we come to the conclusion that n=3 for BSMOTE-PCA and BSMOTE-LDA and d=50 for BSMOTE-PCA will be a better choice.

5. Experimental results and analysis

Our results are obtained by analysing the ROC curves and calculate the AUC. The ROC curve refers to the receiver operating characteristic curve, which is a comprehensive index reflecting the sensitivity and specificity of continuous variables. It is a method of mapping the relationship between

sensitivity and specificity. It is determined by setting the continuous variable $(1-specific)$ as the abscissa plotted into a curve. On the ROC curve, the point closest to the top left of the graph is the critical value for which the sensitivity and specificity are high.

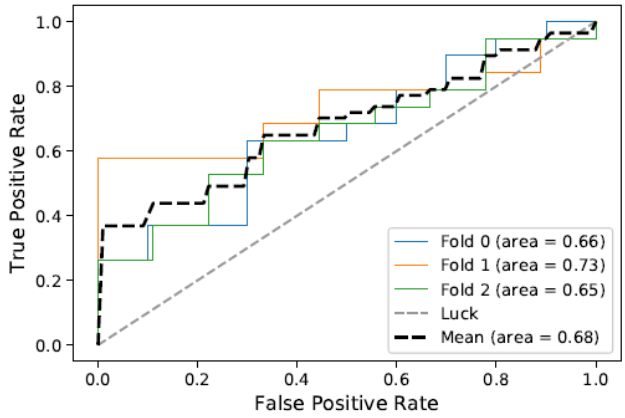
Table 1 Experiment detail for $Data_1$

Fold	Only-SVM			BSMOTE-PCA50			BSMOTE-PCA1			BSMOTE-LDA1		
	Sen	Spe	AUC	Sen	Spe	AUC	Sen	Spe	AUC	Sen	Spe	AUC
0	0.63	0.70	0.66	0.79	0.60	0.69	0.74	0.60	0.56	0.89	0.50	0.77
1	0.58	1.00	0.73	0.58	1.00	0.74	0.42	0.78	0.54	0.74	0.78	0.75
2	0.95	0.22	0.65	0.68	0.67	0.67	0.79	0.44	0.57	0.79	0.89	0.89
Avg	0.72	0.64	0.68	0.68	0.76	0.70	0.65	0.61	0.56	0.81	0.72	0.80

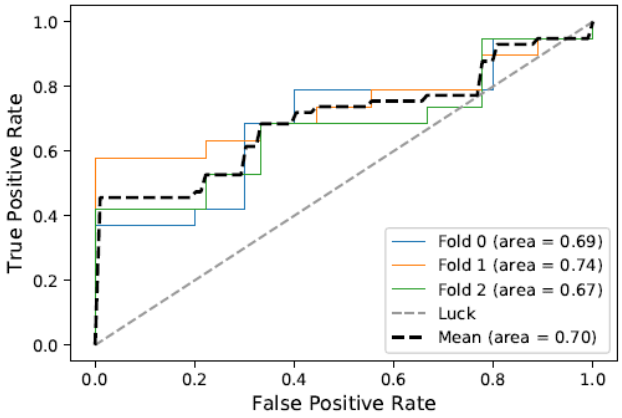
[1]Sen, Spe and Avg mean sensitivity ,specificity and average, respectively.

We calculate the AUC for $Data_1$ and $Data_2$ with 3-fold and 5-fold cross-validation because the samples numbers of these data set are different. $Data_1$ has more samples than $Data_2$ so we need enough samples for training in experiment of $Data_2$.

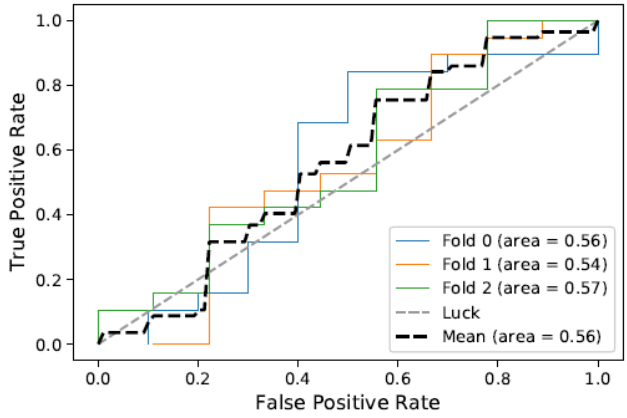
Figure 4 shows the performance of methods in experiment of $Data_1$. Compared with the methods of Only-SVM, BSMOTE-PCA50 and BSMOTE-PCA1, our method has better performance. Only-SVM seems to be sensitive to class-imbalance and BSMOTE-PCA50 has similar AUC with Only-SVM because the PCA with 50 principal component reduce small part of the noise in the data but has not extracted important information of data.



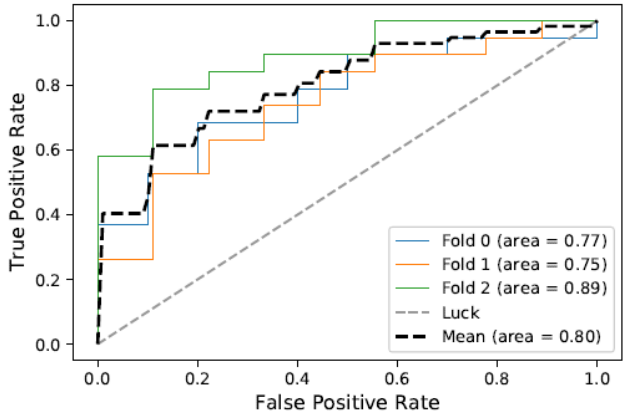
(a) Only-SVM



(b) BSMOTE-PCA50



(c) BSMOTE-PCA1



(d) BSMOTE-LDA1

Fig. 4 ROC graphic of four different models in $Data_1$.

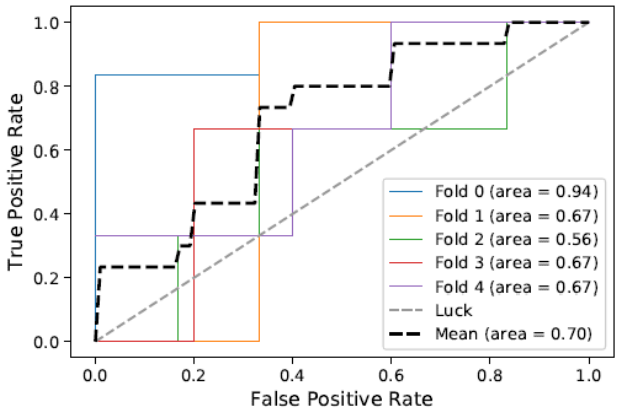
After over-sampling with BSMOTE, the samples number of minority class and majority class are close. However, the BSMOTE-PCA1 perform the worst due to incorrect feature extraction. A lot of information lost and the only feature we get from PCA cannot lead to better classification. We conduct

supervised dimension reduction (using LDA) for our method (the BSMOTE). The only one feature obtained from LDA is actually the weighting of multiple features. Figure 4(d) showed the ROC of BSMOTE-LDA1.

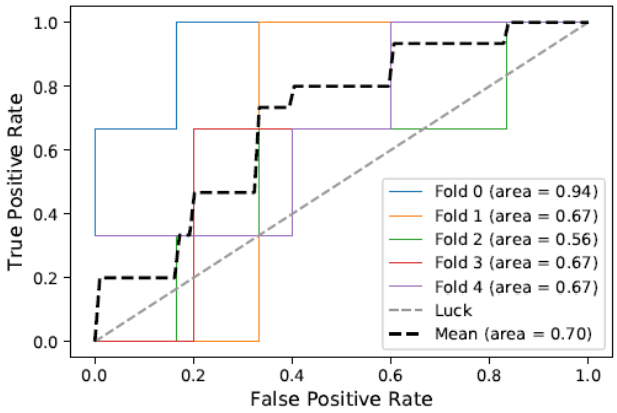
The details of our first experiment results are showed in Table 1. The area under curve of Only-SVM, BSMOTE-PCA50, BSMOTE-PCA1 and BSMOTE-LDA1 models are 68%, 70%, 56% and 80%. Figure 5 and Table 2 show the results of the second experiment and the AUC of each methods are 70%, 70%, 68% and 77%. Different results produce from different data set but BSMOTE-LDA1 still has higher AUC than other methods in our experiment.

Table 2 Experiment detail for $Data_2$

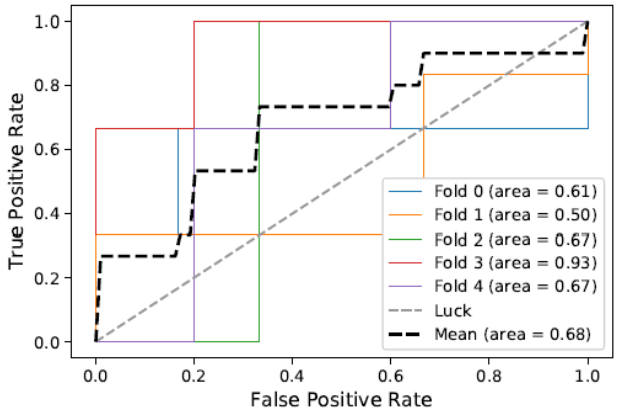
Fold	Only-SVM			BSMOTE-PCA50			BSMOTE-PCA1			BSMOTE-LDA1		
	Sen	Spe	AUC	Sen	Spe	AUC	Sen	Spe	AUC	Sen	Spe	AUC
0	0.83	1.00	0.94	1.00	0.80	0.94	0.67	0.83	0.61	1.00	1.00	1.00
1	0.67	1.00	0.67	1.00	0.67	0.67	0.33	1.00	0.50	1.00	0.83	0.94
2	0.67	0.67	0.56	0.67	0.67	0.56	1.00	0.67	0.67	0.67	0.67	0.56
3	0.67	0.83	0.67	0.67	0.83	0.67	1.00	0.80	0.93	1.00	0.60	0.80
4	1.00	0.40	0.67	1.00	0.40	0.67	0.67	0.80	0.67	0.20	1.00	0.53
Avg	0.77	0.78	0.70	0.87	0.67	0.70	0.73	0.82	0.68	0.77	0.82	0.77



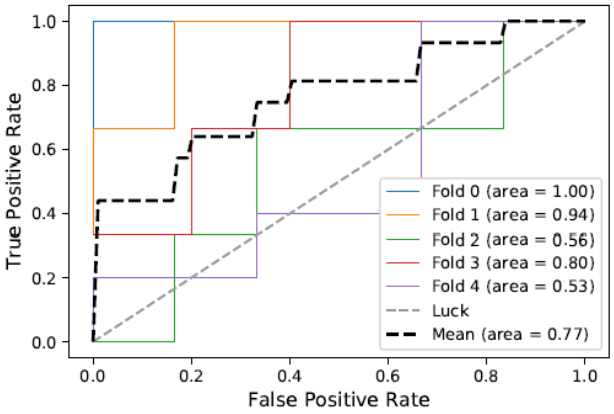
(a) Only-SVM



(b) BSMOTE-PCA50



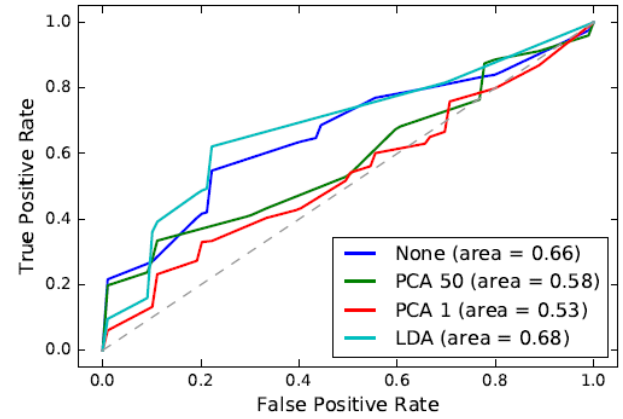
(c) BSMOTE-PCA1



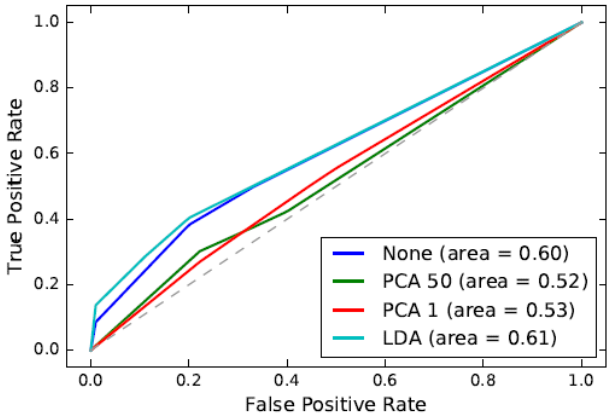
(d) BSMOTE-LDA1

Fig. 5 ROC graphics of four different models in $Data_2$.

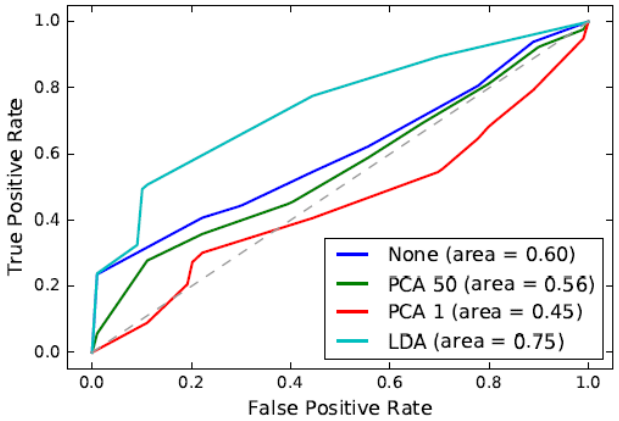
In addition, BSMOTE-LDA1 not only perforce well in SVM but also other three frequently used classifications. Figure 6 shows the performance of our method in different classifications and the details are given in Table 3. From these results we can see the better performance of BSMOTE-LDA1 model.



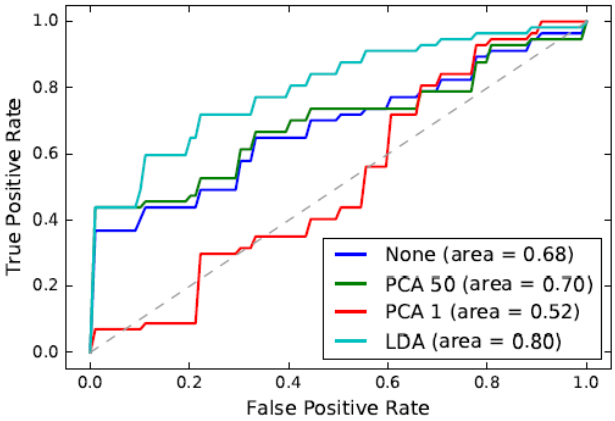
(a) Random Forest



(b) Decision Tree



(c) k-Nearest Neighbor



(d) Support Vector Machines

Fig. 6: Comparison of different methods in different classification are displayed. In this part, samples of $Data_1$ are used to train and test. Specifically, label *None* means samples of test set are classified directly by these four different methods.

Table 3 AUC of comparisons in different classification

	None	BSMOTE-PCA50	BSMOTE-PCA1	BSMOTE-LDA1
RF	0.66	0.58	0.53	0.68
DT	0.60	0.52	0.53	0.61
KNN	0.60	0.56	0.45	0.75
SVM	0.68	0.70	0.52	0.80

6. Conclusion

In recent years, learning with high-dimensional and class-imbalanced data sets has received more and more attentions in both practical and theoretical aspects. However, traditional methods are not satisfactory and we combine BSMOTE with LDA to solve this problem.

BSMOTE is an improved algorithm which only over-sample the minority examples near the borderline. For dimension reduction, LDA is a supervised method and more suitable for classification. To verify the advantage of our method, we conduct experiment to compare BSMOTE-LDA with other methods such as the combination of BSMOTE and PCA. It is demonstrated by experiments that our method has better performance.

References

[1] Pandey, M. Mann, Proteomics to study genes and genomes., Nature 405 (6788) (2000) 837-46.
[2] M. R. Wilkins, J. C. Sanchez, A. A. Gooley, R. D. Appel, I. Humpherysmith, D. F. Hochstrasser, K. L. Williams, Progress with proteome projects: why all proteins expressed by a genome should

be identifed and how to do it., Biotechnology & Genetic Engineering Reviews 13 (1) (1996) 19-50.

[3] R. Matthiesen, Mass spectrometry data analysis in proteomics, Methods in Molecular Biology 1007 (2013) 1064-3745..

[4] Poersch, G. M. Lopes, V. P. de Carvalho, G. P. Lanfredi, S. P. C. De, L. J. Greene, C. B. de Sousa, H. H. Angotti Carrara, C. D. R. Fj, V. M. Faca, A proteomic signature of ovarian cancer tumor uid identi ed by highthroughput and veri ed by targeted proteomics, Journal of Proteomics 145 (2016) 226-236.

[4] N. Cruz, H. M. Coley, H. B. Kramer, T. K. Madhuri, N. A. Safuwan, A. R. Angelino, M. Yang, Proteomics analysis of ovarian cancer cell lines and tissues reveals drug resistance-associated proteins, Cancer Genomics & Proteomics 14 (1) (2017) 35.

[5] N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study, Intelligent Data Analysis 6 (5) (2002) 429-449.

[6] [7]N. V. Chawla, N. Japkowicz, Editorial: special issue on learning from im-balanced data sets, SIGKDD Explor. Newsl (2004) 1-6

[7] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, Enrichr: a comprehensive gene set enrichment analysis web server 2016 update, Nu-cleic Acids Research 44 (Web Server issue) (2016) W90-W97.

[8] P. H. Guzzi, Microarray data analysis. methods and applications. second edition, Anticancer research 36 (4) (2016) 2045.

[9] R. Blagus, L. Lusa, Class prediction for high-dimensional class-imbalanced data, Bmc Bioinformatics 11 (1) (2010) 523.

[10] Xie, Z. Qiu, The e ect of imbalanced data sets on lda: A theoretical and empirical analysis, Pattern Recognition 40 (2) (2007) 557-562.

[11] H. Han, W. Y. Wang, B. H. Mao, Borderline-smote: A new over-sampling 345 method in imbalanced data sets learning, in: International Conference on Intelligent Computing, 2005, pp. 878-887.

[12] D. Zhuang, B. Zhang, Q. Yang, J. Yan, Z. Chen, Y. Chen, E cient text classi cation by weighted proximal svm, in: IEEE International Conference on Data Mining, 2005, pp. 538-545.

[13] H. M. Nguyen, E. W. Cooper, K. Kamei, Borderline over-sampling for imbalanced data classi cation, Inderscience Publishers, 2011.

[14] S. Pouyanfar, S. C. Chen, Automatic video event detection for imbalance data using enhanced ensemble deep learning, International Journal of Se-mantic Computing 11 (1) (2017) 85-109.
A. Jesson, N. Guizard, S. H. Ghalehjegh, D. Goblot, F. Soudan, N. Chapa-dos, CASED: Curriculum Adaptive Sampling for Extreme Data Imbalance, 2017.

[15] S. Ando, C.-Y. Huang, Deep Over-sampling Framework for Classifying Im-balanced Data, ArXiv e-printsarXiv:1704.07515.

[16] M. Buda, A. Maki, M. A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, ArXiv e-printsarXiv: 1710.05381.

[17] Z. Wang, Y. Wang, J. Xuan, Y. Dong, M. Bakay, Y. Feng, R. Clarke, E. P. Ho man, Optimized multilayer perceptrons for molecular classi cation and diagnosis using genomic data., Bioinformatics 22 (6) (2006) 755-761.

[18] H. Abdi, L. J. Williams, Principal component analysis, Wiley Interdisci-plinary Reviews Computational Statistics 2 (4) (2010) 433-459.

[19] S. Balakrishnama, A. Ganapathiraju, Linear discriminant analysis|a brief tutorial, Proc.of the Int.joint Conf.on Neural Networks 3 (94) (1998) 387-391.

[20] R. E. Prieto, A general solution to the maximization of the multidimen-sional generalized rayleigh quotient used in linear discriminant analysis for signal classi cation, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings, 2003, pp. VI-157-60 vol.6

[21] G. V. S. George, V. C. Raj, Review on feature selection techniques and the impact of svm for cancer classi cation using gene expression pro le, International Journal of Computer Science & Engineering Survey 02 (03).
[22] Jenkins, W. Nick, K. Roy, A. Esterline, J. Bloch, Author identi cation using sequential minimal optimization, in: Southeastcon, 2016, pp. 1-2.
A. P. Bradley, The use of the area under the roc curve in the evaluation of machine learning algorithms, Pattern Recognition 30 (7) (1997) 1145-1159.
[23] J. A. Hanley, B. J. Mcneil, The meaning and use of the area under a receiver operating characteristic (roc) curve., Radiology 143 (1) (1982) 29.
[24] N. V. Chawla, N. Japkowicz, A. Kotcz, Editorial:special issue on learn-ing from imbalanced data sets, Acm Sigkdd Explorations Newsletter 6 (1) (2004) 1-6.
[25] Breiman, Random forests-random features, Machine Learning 45 (1) (1999) 5-32.
[26] J. R. Quinlan, Induction on decision tree, Machine Learning 1 (1) (1986) 81-106.
[27] T. M. Cover, P. E. Hart, Nearest neighbor pattern classi cation, IEEE Trans.inf.theory 13 (1) (1967) 21-27.