An Improved Ensemble Learning Method for Classifying High-Dimensional and Imbalanced Biomedicine Data

Hualong Yu and Jun Ni

Abstract—Training classifiers on skewed data can be technically challenging tasks, especially if the data is high-dimensional simultaneously, the tasks can become more difficult. In biomedicine field, skewed data type often appears. In this study, we try to deal with this problem by combining asymmetric bagging ensemble classifier (asBagging) that has been presented in previous work and an improved random subspace (RS) generation strategy that is called feature subspace (FSS). Specifically, FSS is a novel method to promote the balance level between accuracy and diversity of base classifiers in asBagging. In view of the strong generalization capability of support vector machine (SVM), we adopt it to be base classifier. Extensive experiments on four benchmark biomedicine data sets indicate that the proposed ensemble learning method outperforms many baseline approaches in terms of *Accuracy*, *F-measure*, *G-mean* and *AUC* evaluation criterions, thus it can be regarded as an effective and efficient tool to deal with high-dimensional and imbalanced biomedical data.

Index Terms—Bioinformatics, class imbalance, ensemble learning, high-dimensional biomedicine data

1 Introduction

In real-world applications, class imbalance problem occurs frequently. The problem usually causes great underestimation for the classification performance of the minority class, further generates unauthentic and meaningless recognition results [1]. For example, in medical diagnosis, the number of patients is generally much smaller than that of healthy people. If a learning algorithm classify all samples into the majority class, it can minimize the error rate. In this case, however, all patients will be misdiagnosed, thus the classifier should be specifically designed when class imbalance problem emerges.

Indeed, class imbalance problem has drawn a significant amount of interests from artificial intelligence, data mining and machine learning in the past decade, reflecting in the installments of several major workshops and special issues, including AAAI'00 [2], ICML'03 [3] and ACM SIGKDD Explorations Newsletter'04 [1]. Lots of solutions have also been presented to address class imbalance problem, including resampling [4], [5], [6], [7], cost sensitive learning [8], [9], [10], ensemble learning [11], [12], [13], [14], one class learning [15], [16], Kernel-based classifier [17], [18] and active learning [19], [20], etc.

However, majority previous work paid more attentions to the development of methods but ignored the influences of the data type. Some recent work [21], [22], [23], [24] found that the classification performance can severely degenerate if directly implementing class imbalance learning methods on high-dimensional, high noise or small sample data sets. When these features emerge simultaneously, the damage can be further intensified. We notice that this data type is ubiquitous in biomedicine field and has wide applications, especially some applications based on Omics data, including cancer diagnosis using DNA microarray data [25], [26]/protein mass-spectrometry data [27], [28], DNA translation initiation detection [29], recognition of microRNA precursors based on DNA sequences data [30], activity prediction adopting drug molecules data [14] etc. It propels us to develop an effective and efficient learning method to classify these high-dimensional and imbalanced biomedical data.

To overcome the problems caused by the high dimensional and imbalanced biomedical data, we investigate the characteristics of this data type and propose a novel and hybrid ensemble learning solution: asBagging_FSS (asymmetric bagging ensemble classifier with feature subspace (FSS)). First, we utilize clustering and feature selection to filter redundant and noisy features, respectively. By both technologies, we can successfully transform training samples from high-dimensional space to low-dimensional feature space. Then we apply a random project function to generate multiple so-called feature subspaces, further adopt bootstrap strategy to extract the same number of majority class samples with that of the minority class in each subspace. Note that FSS is one improved version of random subspace (RS) and it can increase accuracy of each base classifier in asBagging with the minimal sacrifice of diversity. Finally, we deploy support vector machine (SVM) [31] on each training subset to generate the final decisions for those unseen test instances by majority voting. We tested the proposed hybrid method

- *H.L. Yu is with the School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang, 212003, China. E-mail: yuhualong@just.edu.cn.*
- *J. Ni is with the Department of Radiology, Carver College of Medicine, The University of Iowa, Iowa City, IA 52242. E-mail: jun-ni@uiowa.edu.*

Manuscript received 29 Dec. 2013; accepted 24 Jan. 2014. Date of publication 23 Feb. 2014; date of current version 4 Aug. 2014.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2014.2306838

on four baseline biomedical data sets and acquired promising results.

The rest of this paper is organized as follows. Section 2 simply reviews some previous work related to class imbalance learning in biomedicine field. In Section 3, we first introduce the basic ideas of asBagging and the proposed FSS generation strategy, then give theoretical analysis for the efficiency of FSS and explain why FSS can help improve the classification performance of asBagging to some extent, finally describe the process of the hybrid ensemble learning method in detail. Section 4 presents the experimental results and discussions based on these results. At last, we conclude this paper and indicate future research in Section 5.

2 Related Work

In past several years, with rapid development of high throughput detection technologies, we are provided more opportunities to develop new applications in biomedicine field. We observe that most available biomedicine data is high-dimensional, including DNA microarray data [25], [26], protein mass-spectrometry data [27], [28], nucleic acid data [29], amino acid sequence data [32], drug molecules activity data [14] etc. Meanwhile, the practical applications often request imbalanced classification, e.g., in DNA microarray data and protein mass-spectrometry data, the number of cancer instances are usually smaller than that of healthy people; microRNAs appear sparsely in DNA sequences; drug targets take only a fraction of drug molecules activity data. However, in these cases, the minority class instances often contain more significant biology or medicine meanings, thus researching how to promote the classification performance of high dimensional and skewed data has important meanings in biomedicine field.

Recent work has focused on class imbalance problem in biomedicine field and has developed several effective methods. One frequently used method is to divide the original data set into a balanced data set for training and an imbalanced data set for testing. The method was used to diagnose myocardial perfusion by cardiac Single Proton Emission Computed Tomography (SPECT) images [33] and to predict polyadenylation signals in human sequences [34]. We note that the possible loss of lots of potentially useful majority class instances in this method usually makes the classifier useless.

Resampling is another popular solution for class imbalance problem. Kamal et al. [25] made use of random oversampling (ROS) to balance gene expression data, further guided feature gene selection and cancer classification. Batuwita and Palade [30] integrated SMOTE oversampling strategy [4] into SVM classifier to predict human microRNA genes. Dobson et al. [35] combined random undersampling (RUS) technology and decision tree classifier to discriminate a few deleterious nsSNPs from a mass of neutral nsSNPs. In our recent work, we proposed an ant colony optimization-based undersampling approach to improve the recognition rate of cancer gene expression samples [26]. Resampling is both easy and intuitionistic, but the promotion of classification performance is quite limited.

The most popular class imbalance learning method is to combine resampling technology together with an ensemble classifier. Ozcift combined data resampling and random forest classifier to diagnose cardiac arrhythmia with acquiring an improved recognition rate [36]. Oh et al. [20] proposed an ensemble learning model with active sample selection to classify biomedical data. The most direct solution is asymmetric bagging ensemble classifier (asBagging) which was initially applied to retrieve images [13]. asBagging generates multiple diverse balanced training subsets by bootstrapping majority class examples with replacement. asBagging and its similar version have been used to predict activity of drug molecules data [14] and to classify proteins [32]. In contrast with the other ensemble learning approaches, we observe that asBagging often consumes less training time with higher or at least comparable classification performance. However, we also find that it is difficult to gurantee the diversity of base classifiers in asBagging merely by bootstrap technology. Furthermore, considering high dimension can often destroy the performance of class imbalance learning methods [21], [22], we try to design an improved asBagging classifier that is specialized for classifying high dimensional and skewed biomedicine data in this study.

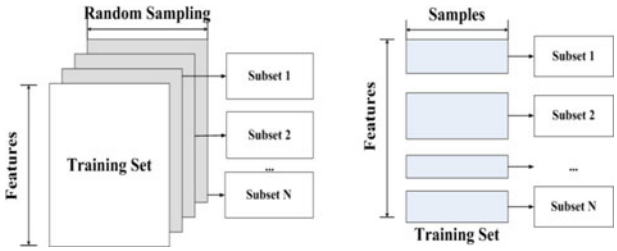


Fig. 1. Graphical representations of Bagging and RS. Left subgraph: Bagging; Right subgraph: RS.

3 Methods

3.1 Asymmetric Bagging Classifier

Bagging, as one of the most important and successful ensemble learning models, incorporates bootstrap and aggregation [37], where bootstrap is a statistics approach using random sampling, i.e., randomly extracting training examples with replacement. In Bagging, multiple base classifiers can be generated upon some randomly extracted training subsets. Generated classifiers are called *weak classifiers* and they can be aggregated to be a *strong classifier* by majority voting rule. Similar to Bagging, random subspace [38] also integrates both bootstrap and aggregation. But RS is different from Bagging due to it runs Bootstrap in feature space. Generally speaking, RS performs better on high dimensional data sets, because in this scenario, acquiring diverse training subsets can become much easier. The schemas of Bagging and RS are described in Fig. 1.

Since the direct use of Bagging or RS is not appropriate for imbalanced classification tasks, asymmetric bagging (asBagging) ensemble classifier [13], [14] was proposed in previous work. For asBagging, bootstrap is only executed on majority class samples, while scarce minority class examples can be completely reserved. This can be seen as an approximate random undersampling procedure on each training subset. Considering the superiority of RS on

high-dimensional data, it has also been integrated into asBagging learning model [13]. For another puzzled problem, i.e., why asBagging works? Tao et al. [13] has provided detailed theoretical analysis.

3.2 Feature Subspace

Much previous work has indicated that a sufficient and necessary condition the ensemble classifier outperforms its individual members is that base classifiers should be simultaneously accurate and diverse [39]:

$$E = \overline{E} - \overline{A}, \tag{1}$$

where E is ensemble generalization error, $\overline{E}$ and $\overline{A}$ are average of generalization errors and diversities of all base classifiers, respectively. To produce successful ensemble learning models, two aspects should be considered simultaneously.

It is clear that for high-dimensional classification tasks with lots of noisy and redundant features, neither Bagging nor RS performs well, thus we design feature subspace generation strategy to alleviate this problem in this study. As one variant of RS, FSS combines both clustering and feature selection, where clustering is adopted to filter redundant features and feature selection is used to eliminate noisy features. In contrast with RS, FSS is expected to improve accuracy of each base classifier with small loss of diversity.

To construct FSS, we first need to extract feature space which can be seen as one feature subset of original feature space. We expect that the irrelevant and redundant features can be removed in this condensed space. In order to delete redundant features, similar features need to be collected into multiple different groups by hierarchical clustering method that utlizes Pearson correlation coefficient as similarity measurement [40]. Pearson correlation coefficient computes the similarity between two features f_i and f_j with

$$Sim(f_i, f_j) = \frac{\sum_{k=1}^{N}(f_{ik} - \overline{f_i})(f_{jk} - \overline{f_j})}{\sqrt{\sum_{k=1}^{N}(f_{ik} - \overline{f_i})^2}\sqrt{\sum_{k=1}^{N}(f_{jk} - \overline{f_j})^2}}, \tag{2}$$

where f_{ik} is the value of f_i on the kth sample, $\overline{f_i}$ represents the mean of f_i, N denotes the number of training samples and $Sim(f_i, f_j)$ denotes the correlation strength between f_i and f_j. It is obvious that the larger the Pearson correlation coefficient of two features is, the more similar they are. After acquiring multiple clusters, the most relative feature of the classification task in each cluster is extracted. In this study, we use signal-noise ratio (SNR) [41] to extract these features. The computational formula of SNR is described as follows:

$$SNR(f_i) = |\mu_0 - \mu_1|/(\sigma_0 + \sigma_1), \tag{3}$$

where μ_0 and μ_1 are mean values of feature f_i belonging to two different classes, σ_0 and σ_1 are their standard deviations, respectively. It is clear that the extracted features are closely related with classification and they are approximatively non-redundant. The set containing only these extracted features is defined as feature space. Suppose the dimension of feature space is d and that of feature subspace is k ($k \leq d$), respectively. Then one random project function P can estabilish the mapping relationship between d and k

$$P(R^d) \in R^k. \tag{4}$$

Making use of the random project function P, it is easy to construct multiple diverse feature subspaces. Predictably, the generated base classifiers by different feature subspaces can be more accurate than those in RS and more diverse than those in Bagging. In other words, it is expected that FSS can improve the balance level between accuracy and diversity of base classifiers. The pseudo-code and schematic illustration of the proposed FSS generation algorithm are described in Figs. 2 and 3, respectively.

Input: Training set S; Feature set F; Weak classifier I; Size of feature space d; Size of feature subspace k; Size of ensemble T; Testing samples x.
Process:
1. Gather features of F into d clusters by hierarchical clustering: CLUSTER$_i$ ($1 \leq i \leq d$);
2. For i=1:d
3. {
4. Select representative feature f_i in CLUSTER$_i$ by SNR;
5. }
6. Construct feature space FS including all representive features extracted above;
7. For i=1 to T
8. {
9. FSS$_i$ =P(FS$\in R^d$)$\in R^k$; /* FSS: feature subspace
10. C_i=I(FSS$_i$, S);
11. }
12. $C^*(x)$=aggregation$\{C_i(x), 1 \leq i \leq T\}$
Output: Ensemble classifier C^*.

Fig. 2. Pseudo-code description for FSS generation algorithm.

3.3 Theoretical Analysis for the Efficiency of FSS

In this section, we give theoretical analysis for the efficiency of FSS. Suppose f is one feature in feature space and we have put it into the feature subspace S_i, then the probability of finding the same feature f in another feature subspace S_j can be calculated by the following formula:

$$P(f \in S_j | f \in S_i) = k/d. \tag{5}$$

That means for two randomly generated feature subspaces, their coselection rate is k/d in theory. Meanwhile, because two different features in feature space can be seen as approximately non-redundant, the theoretical diversity div between two feature subspaces can be calculated by

$$div = (d - k)/d. \tag{6}$$

Obviously, when $d \gg k$, the diversity of feature subspaces can be guaranteed securely. Furthermore, we observe that for k dimension feature subspace, the number of diverse combinations is

$$C_d^k = \frac{d!}{(d-k)!k!}, \tag{7}$$

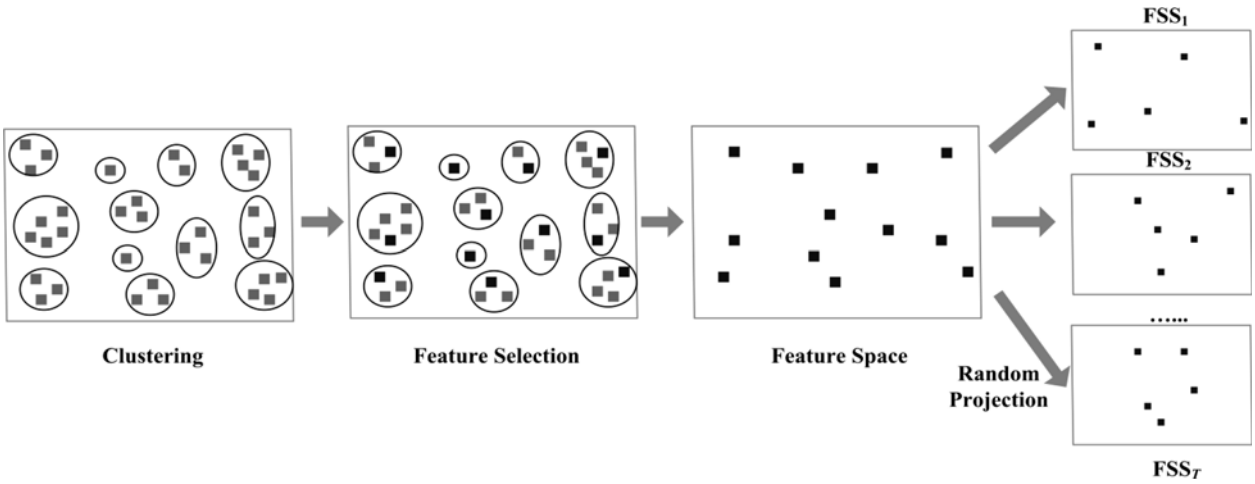


Fig. 3. Schematic illustration of FSS generation algorithm.

i.e., when there is a obvious difference between d and k, abundant diverse base learners can be generated to construct an excellent enough ensemble learning model.

In contrast with Bagging, it is not difficult to find that FSS is more diverse and meanwhile the loss of accuracy is relatively small due to in each feature subspace, all features are strongly related to classification task. While compared with RS, FSS is more accurate with a few sacrifice of diversity. In other words, FSS improves the balance level between accuracy and diversity of base classifiers, further helps improve classification performance to some extent.

3.4 asBagging_FSS Classifier

asBagging_FSS aims to promote the performance of imbalanced classification tasks by combining asBagging and FSS. Fig. 4 gives the detailed pseudo-code description of asBagging_FSS.

Fig. 5 schematically illustrates the workflow of asBagging_FSS ensemble learning model. It can be seen that the model comprises five components as follows: 1) single feature space is generated by both clustering and feature selection; 2) a mass of FSSs are extracted randomly from the feature space by the random project function; 3) some balanced training subsets are built upon the corresponding FSSs by Bootstrap; 4) multiple base classifiers are generated based on the corresponding training subsets and 5) a voting rule is created upon the results of all base classifiers to make the final decision.

In these components, we pay particular attention for the type of baseline classifier. Considering the solid theory foundation and specific design for high dimensional and small sample data, support vector machine [31] is adopted as base classifier in our ensemble learning model. Another interesting point is the determination of the size of feature subspace which can seriously affect the final classification performance. In next section, we will tune this parameter gradually and evaluate its influence.

Moreover, we apply the simplest voting rule, i.e., majority voting, for the final decision-making. Its merits lie in

Input: Minority class training set S^+; Majority class training set S^-; Feature set F; Weak classifier I; Size of feature space d; Size of feature subspace k; Size of ensemble T; Testing samples x.
Process:
1. Gather features of F into d clusters by hierarchical clustering: $CLUSTER_i$ ($1 \le i \le d$);
2. For i=1:d
3. {
4. Select representative feature f_i in $CLUSTER_i$ by SNR;
5. }
6. Construct feature space FS including all representive features extracted above;
7. For i=1 to T
8. {
9. $FSS_i = P(FS \in R^d) \in R^k$; /* FSS: feature subspace
10. S_i^- =Bootstrap examples from S^-, s.t. $|S^+| = |S_i^-|$;
11. $C_i = I(FSS_i, S_i^-, S^+)$;
12. }
13. $C^*(x)$=aggregation$\{C_i(x), 1 \le i \le T\}$
Output: Ensemble classifier C^*.

Fig. 4. Pseudo-code for asBagging_FSS algorithm.

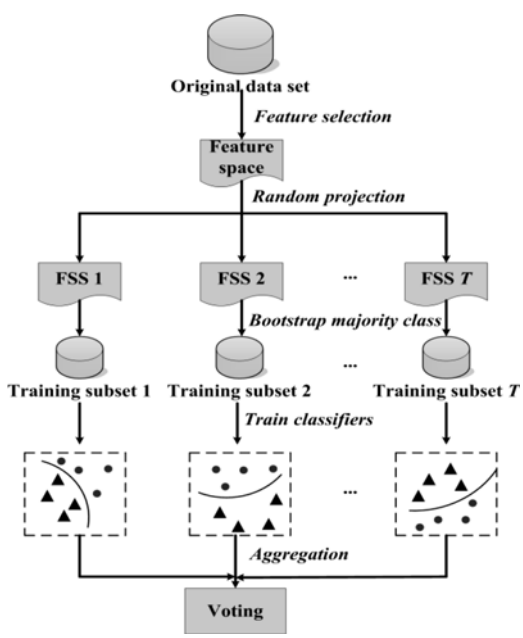


Fig. 5. Schematic illustration of asBagging_FSS algorithm.

TABLE 1
Biomedical Data Sets Used in This Study

Data Set	Number of samples	Number of features	Imbalance Ratio
Colon	62 (22:40)	2000	1.82
Lung	39 (15:24)	2880	1.60
Ovarian I	116 (16:100)	15154	6.25
Ovarian II	116 (16:100)	15154	6.25

neither requiring any priori knowledge nor requiring any complex and intensive computation [42].

4 Experiments

4.1 Data Preparation

We perform empirical experiments on four real-world high dimensional and imbalanced biomedicine data sets. Two of them are cancer DNA microarray data sets: Colon and Lung [43], [44], two others are cancer protein mass spectrometry data sets: Ovarian I and Ovarian II [45].

Colon data set contains 62 samples collected from colon cancer patients [43]. Among them, 40 tumor biopsies are from tumors and 22 normal biopsies are from healthy parts of the colons of the same patients. 2,000 out of around 6,500 genes were selected based on the confidence in the measured expression levels.

Lung data set includes gene expression data on tumor specimens from a total of 39 non-small cell lung cancer samples [44]. Among these samples, 24 patients had experienced relapse of their tumor either locally or as a distant metastasis. The remaining 15 patients are disease free based on both clinical and radiological testing. The data are described by 2,880 genes.

Ovarian I and Ovarian II data sets consist of 116 mass spectrometry instances derived from serum of women, respectively [45]. The task of Ovarian I is to distinguish 16 benign samples from 100 Ovarian cancer examples, as well as Ovarian II is used to distinguish the same 16 benign samples from 100 unaffected examples. Each sample has 15,154 features.

The detailed information about these four data sets is summarized in Table 1. Colon and Lung data sets can be downloaded from http://datam.i2r.a-star.edu.sg/datasets/krbd/. Ovarian I and Ovarian II data sets are available at: http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp.

4.2 Evaluation Criteria

Generally, in skewed classification tasks, the minority class is labeled as positive and the majority class is marked as negative. Table 2 gives the confusion matrix of a two-class problem, where *TP* and *TN* denote the number of correct positive and negative examples, *FN* and *FP* represent the number of misclassified positive and negative examples, respectively. It is well-known that in skewed data, overall accuracy (*Acc*) often gives bias evaluation, thus some other specific evaluation metrics as true positive rate (*TPR*), true negative rate (*TNR*), *Fmeasure*, *G-mean* and area under the receiver operating characteristic curve (*AUC*), are used as supplementary evaluation criteria. Based on Table 2, these evaluation criteria are calculated as follows:

TABLE 2
Confusion Matrix

	Predicted positive class	Predicted negative class
Actual positive class	*TP* (True Positive)	*FN* (False Negative)
Actual negative class	*FP* (False Positive)	*TN* (True Negative)

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{8}$$

$$TPR = Recall = \frac{TP}{TP + FN}, \tag{9}$$

$$TNR = \frac{TN}{TN + FP}, \tag{10}$$

$$Precision = \frac{TP}{TP + FP}, \tag{11}$$

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{12}$$

$$G\text{-}mean = (TPR \times TNR)^{1/2}. \tag{13}$$

AUC is the area below the ROC curve that depicts the performance of a classifier using the (*FPR*, *TPR*) pairs. It has been proved to be a reliable performance measure for class imbalance problem [46].

4.3 Results and Discussions

During the experiments, we performed 10 times' three-fold cross validation and output all performance metrics in the form of mean±variance. To avoid giving inauthentic classification results, feature selection only relies on the training set in the procedure of cross-validation. Gaussian radial basis function-based SVM was used as base classifier because of its high reliability [47]. To present the superiority of the proposed asBagging_FSS ensemble learning algorithm, we evaluate it in comparison with that of eight other classification approaches, including 1. SVM upon FS (SVM_FS), 2. SVM upon RS (SVM_RS), 3. SVM with RUS (SVM_RUS), 4. ensemble of FS (Bagging_FS), 5. ensemble of RS (Bagging_RS), 6. ensemble of FSS (Bagging_FSS), 7. asymmetric Bagging ensemble of FS (asBagging_FS) and 8. asymmetric Bagging ensemble of RS (asBagging_RS).

For each classification approach, the parameter settings are the same: the parameter σ of RBF function and the penalty factor C in SVM were empirically designated as 10 and 500, respectively [26]. For all ensemble classifiers, T as the amount of base classifiers was designated as 100. The dimension of feature space and feature subspace (or random subspace) were initially designated as 100 and 20,

TABLE 3
Performance Evaluation of Various Methods on Colon Data Set

Method	Performance metrics (mean±variance, %)					
	Acc	*TPR*	*TNR*	*F-measure*	*G-mean*	*AUC*
SVM_FS	80.49±2.33	71.36±5.40	85.50±3.50	72.14±3.44	78.01±2.84	83.92±1.62
SVM_RS	81.94±2.77	73.64±4.90	86.50±4.21	74.33±3.47	79.72±2.79	85.12±2.13
SVM_RUS	82.26±1.77	81.82±7.33	82.50±3.87	76.49±2.98	81.98±2.62	86.06±2.48
Bagging_FS	82.90±2.19	72.27±4.29	**88.75±3.40**	75.01±2.91	80.02±2.30	87.39±1.67
Bagging_RS	80.65±2.88	69.55±7.06	86.75±2.75	71.70±4.82	77.56±4.01	86.02±2.74
Bagging_FSS	83.71±2.44	75.91±6.44	88.00±2.92	76.69±3.87	81.62±3.33	87.13±1.38
asBagging_FS	**84.68±3.32**	84.54±3.64	84.75±5.06	79.76±3.66	84.58±2.94	87.46±1.70
asBagging_RS	78.23±1.49	**87.73±2.09**	73.00±3.32	74.12±0.96	79.98±1.04	86.16±1.96
asBagging_FSS	**84.68±2.31**	87.27±2.73	83.25±3.36	**80.21±2.65**	**85.21±2.08**	**88.14±1.86**

TABLE 4
Performance Evaluation of Various Methods on Lung Data Set

Method	Performance metrics (mean±variance, %)					
	Acc	*TPR*	*TNR*	*F-measure*	*G-mean*	*AUC*
SVM_FS	69.74±3.20	52.00±6.53	80.83±4.25	56.81±5.04	64.66±4.02	75.03±4.57
SVM_RS	70.26±4.76	58.67±7.18	77.50±6.77	60.26±6.13	67.22±4.87	70.25±2.64
SVM_RUS	71.54±4.65	66.67±7.30	74.58±5.42	64.28±5.93	70.38±4.89	77.25±3.58
Bagging_FS	72.31±4.97	58.00±6.00	**81.25±7.74**	61.79±5.43	68.44±4.58	76.10±4.00
Bagging_RS	71.02±4.14	60.00±7.89	77.92±5.29	61.31±5.89	68.17±4.76	70.53±3.95
Bagging_FSS	**72.82±3.28**	59.33±4.67	**81.25±5.35**	62.70±3.78	69.31±3.07	76.85±2.56
asBagging_FS	71.79±5.13	73.33±7.89	70.83±5.27	66.63±6.05	71.97±5.39	78.46±3.95
asBagging_RS	68.72±3.94	78.00±4.27	62.92±5.73	65.79±3.58	69.96±3.80	76.40±1.73
asBagging_FSS	**72.82±3.08**	**79.33±4.67**	68.75±5.97	**69.22±2.57**	**73.70±2.72**	**78.88±2.91**

TABLE 5
Performance Evaluation of Various Methods on Ovarian I Data Set

Method	Performance metrics (mean±variance, %)					
	Acc	*TPR*	*TNR*	*F-measure*	*G-mean*	*AUC*
SVM_FS	90.95±1.40	37.50±9.27	99.50±0.67	52.76±9.65	60.62±7.46	**96.33±1.03**
SVM_RS	87.76±2.34	23.75±11.11	98.00±1.67	34.22±13.88	46.78±11.96	81.15±6.21
SVM_RUS	91.38±2.15	71.25±9.35	94.60±2.06	69.49±7.34	81.91±5.54	94.87±1.87
Bagging_FS	91.72±0.96	41.88±6.28	**99.70±0.46**	58.02±6.22	64.43±4.79	81.50±4.10
Bagging_RS	87.84±1.87	27.50±13.46	97.50±1.28	36.86±15.24	49.86±13.79	85.96±2.64
Bagging_FSS	91.04±0.96	42.50±8.75	98.80±0.60	56.15±7.24	64.46±6.38	96.11±1.48
asBagging_FS	**92.76±1.29**	75.62±5.90	95.50±1.36	74.25±4.18	84.91±3.22	93.30±2.49
asBagging_RS	82.59±2.37	76.88±8.41	83.50±1.96	54.93±5.95	80.00±4.98	87.65±3.47
asBagging_FSS	**92.76±0.88**	**78.13±5.04**	95.10±1.04	**74.83±2.93**	**86.14±2.63**	95.88±0.98

respectively. Tables 3, 4, 5, and 6 report the results of various classification methods on four biomedicine data sets.

From the results in these tables, we observe as follows:

1. Ensemble learning generally outperforms the corresponding single classifier, except for RS series algorithms. In contrast with SVM_RS, Bagging_RS performed a little better on three data sets: Lung, Ovarian I and Ovarian II. However, asBagging_RS acquired the lowest *Acc* values on all data sets, indicating it is unacceptable to combine asBagging and RS.
2. Generally speaking, FSS is superior to FS. In Bagging ensemble learning model, FSS beat FS on three data sets and was merely inferior to FS on Ovarian I data set. While in asBagging ensemble learning model, FSS almost won FS in terms of all evaluation metrics on each data set. Though FSS can be only regarded as a subset of FS, it yields more diverse base classifiers with quite small sacrifice of accuracy.
3. *TPR* and *TNR* are negatively correlated with each other, i.e., *TPR* increases with necessary decline of *TNR* and vice-versa. In skewed biomedicine classification tasks, we often pay more attention to the patterns of positive examples (minority class samples) that are closely related to diseases or POIs, thus increasing *TPR* is our primary goal. Since the adoption of undersampling strategy, SVM_RUS and the asBagging series classifiers can promote *TPR* metric greatly, especially on those highly skewed data sets, such as Ovarian I and Ovarian II.
4. asBagging_FSS is most effective for high-dimensional and imbalanced classification tasks. In

TABLE 6
Performance Evaluation of Various Methods on Ovarian II Data Set

Method	Performance metrics (mean±variance, %)					
	Acc	*TPR*	*TNR*	*F-measure*	*G-mean*	*AUC*
SVM_FS	89.74±2.23	35.00±10.53	98.50±1.50	48.01±12.29	58.06±9.00	94.32±2.21
SVM_RS	86.03±1.14	24.37±7.63	95.90±0.83	32.05±8.48	47.69±7.87	80.76±4.72
SVM_RUS	**90.26±1.09**	76.88±8.41	92.40±1.80	**68.40±3.76**	84.11±4.31	93.46±1.87
Bagging_FS	89.57±0.90	32.50±6.12	**98.70±0.64**	45.94±6.49	58.36±5.29	80.57±5.44
Bagging_RS	87.50±1.04	20.00±10.00	98.30±0.90	29.39±11.96	42.87±11.05	85.31±3.60
Bagging_FSS	90.09±0.97	38.12±8.13	98.40±0.66	51.04±7.03	60.92±6.10	92.85±1.36
asBagging_FS	89.66±1.39	73.75±6.73	92.20±1.08	66.24±4.72	82.38±3.83	89.13±2.58
asBagging_RS	73.19±2.09	**78.13±7.53**	72.40±2.62	44.54±3.15	75.08±3.36	84.50±3.54
asBagging_FSS	89.92±0.68	**78.13±3.13**	91.80±0.87	68.12±1.77	**84.67±1.55**	**94.39±0.95**

terms of three important evaluation metrics, asBagging_FSS output the highest *G-mean* values on all data sets, as well as took the first place on three data sets for *F-measure* and *AUC* evaluation criteria, respectively. asBagging_FS was superior to SVM_RUS on most data sets except Ovarian II, indicating the necessity of integrating ensemble learning and resampling strategies to deal with imbalanced classification tasks.

It is also worth noting that the classification performance of our proposed algorithms can be restricted by many factors, including the size of feature space, the size of feature subspace, the number of base classifiers and several parameters in SVM. Indeed, the size of feature subspace is the most significant factor among them. To make clear its influence mechanism, we design a group of new experiments. Let the dimension of feature subspace vary from 5, 10, 20, 30, 50 to 80, and the other parameters follow the initial settings. The average classification results of 10 random runs are summarized in Fig. 6.

There are some fluctuations, however, Fig. 6 still reveals a common trend that the optimal performance often appears when selecting 20-30 dimensions' feature subspace. With further increase of dimension of feature subspace, the classification performance drops rapidly. That means selecting 20-30 dimensions' feature subspace can maximize the balance relationship between accuracy and diversity of base classifiers. It is not difficult to analyze the reason: extracting too few features can damage the accuracy of each base classifier, yet using too many features can hurt the diversity of base classifiers. In fact, in practical applications, the optimal dimension can be determined by internal multiple-fold cross validation inside the training sets. The experimental results help guide us to construct the optimal classification model.

4.4 Why asBagging_FSS Beats Others?

According to the experimental results above, we find that asBagging_FSS can help improve performance of high dimensional and skewed classification tasks and beat many

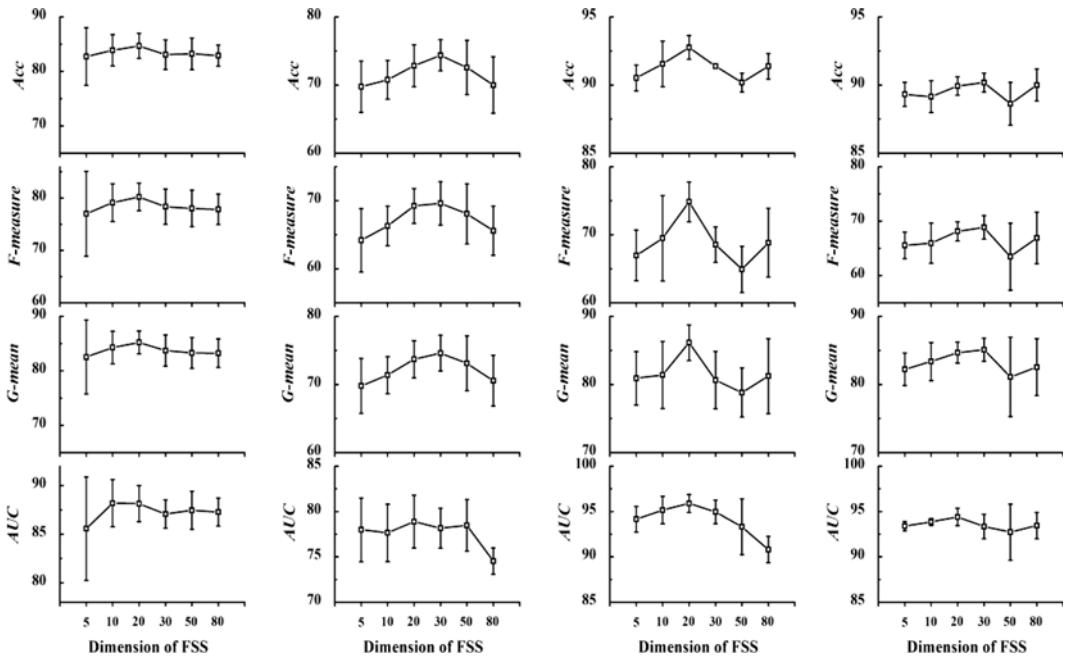
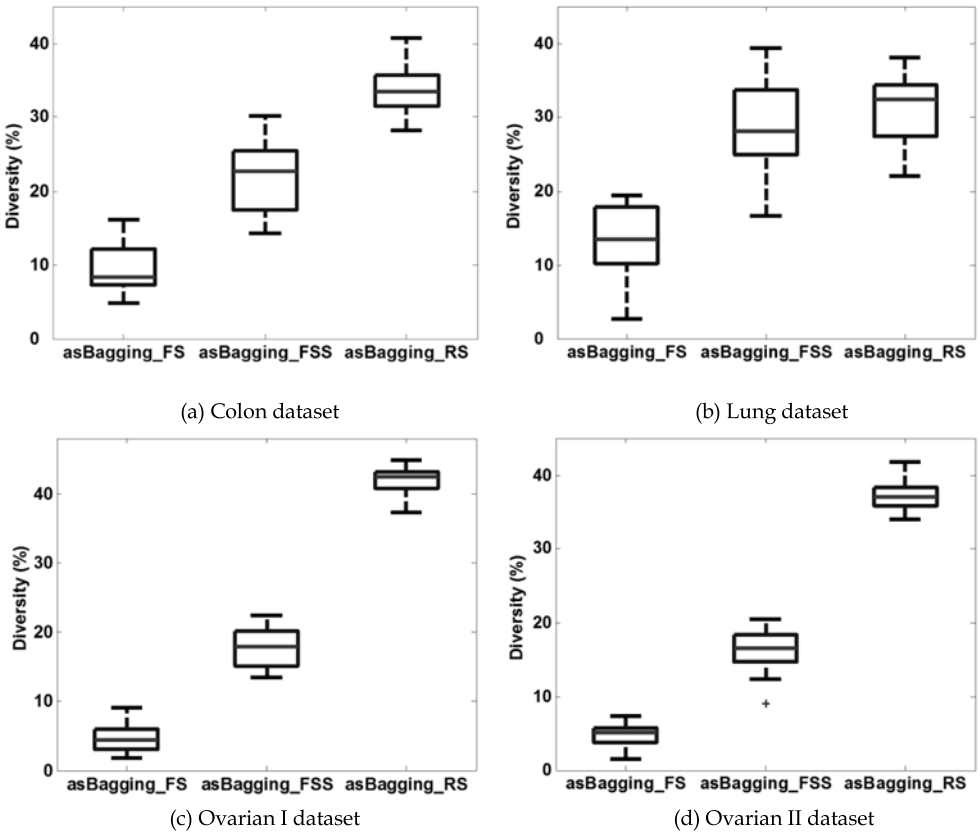


Fig. 6. Performance comparison for asBagging_FSS based on different dimensions of feature subspace. *1*st column: Colon data set; *2*nd column: Lung data set; *3*rd column: Ovarian I data set; *4*th column: Ovarian II data set.



(a) Colon dataset (b) Lung dataset

(c) Ovarian I dataset (d) Ovarian II dataset

Fig. 7. Comparison of average diversity for asBagging_FS, asBagging_FSS and asBagging_RS ensemble learning methods.

traditional classifiers. Then a question appears: why asBagging_FSS beats others? To find the answer, we evaluate the average diversity and average accuracy of the base classifiers in asBagging_FS, asBagging_FSS and asBagging_RS, respectively (see Figs. 7 and 8). The average diversity is detected by using disagreement measure [48]. Disagreement measure first runs each base classifier on test set and acquires the corresponding binary prediction sequence, then compares the difference between every two sequences by counting their mismatched bits. Indeed, the average percentage of mismatched bits can be seen as their diversity. The average accuracy can be detected in a similar way.

From Figs. 7 and 8, we observe that the diversity and the accuracy are negatively correlated with each other, i.e., one of them increases with inevitable decline of the other. asBagging_FS has the smallest diversity but the highest accuracy that is totally contrary to asBagging_RS. As a compromising choice, the proposed asBagging_FSS can greatly enhance the diversity of base classifiers without obvious sacrifice of accuracy in comparison to asBagging_FS. This explains why asBagging_FSS can be superior to those baseline classification algorithms.

Note that though the proposed classifier usually performs well for imbalanced classification tasks, it is not suitable for processing low-dimensional data. Because only for high-dimensional data, FSS can help yield enough diverse training subsets and promote the efficiency of ensemble learning.

Of course, the time complexity of asBagging_FSS is relatively higher than its predecessors. However, this increment is slight and can even be ignored in practical applications, so we didn't compare the running time of various methods in this Section.

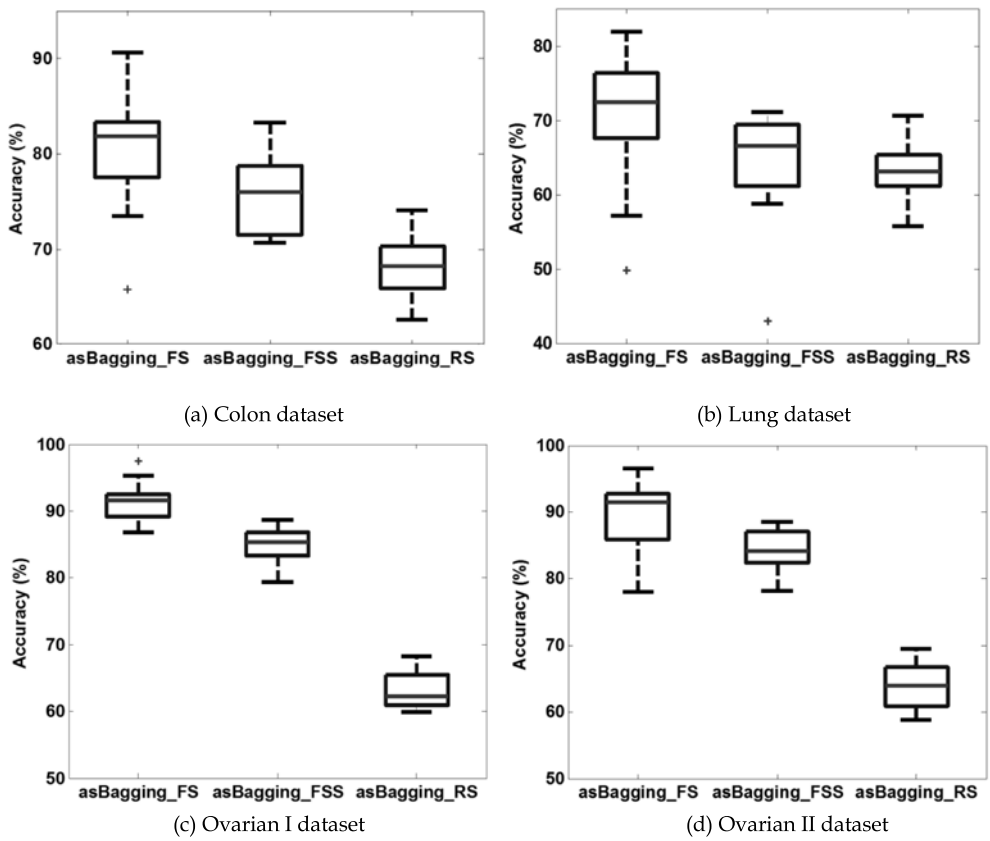
5 Conclusion and Future Work

In this article, one novel ensemble learning approach named as asBagging_FSS is proposed to deal with high dimensional and imbalanced biomedical classification tasks. As one variant of random subspace, the proposed FSS strategy can improve balance level between diversity and accuracy of base classifiers by extracting random feature subspaces which only contain those features which are strongly related with classification tasks and non-redundant with each other. To verify its effectiveness, we tested it on four real high-dimensional and skewed biomedicine data sets and acquired promising results.

We wish our asBagging_FSS algorithm can be applied in some real-world biomedical applications with high dimensional and skewed data, in the near future. Additionally, we will investigate the possibility of extending current asBagging_FSS algorithm to multiclass tasks in future work, too.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grant No.61305058, Natural Science Foundation of Jiangsu Province of China under Grant No. BK20130471, Nature Science Foundation of the Jiangsu Higher Education Institutes of China under grant



(a) Colon dataset (b) Lung dataset

(c) Ovarian I dataset (d) Ovarian II dataset

Fig. 8. Comparison of average accuracy for asBagging_FS, asBagging_FSS and asBagging_RS ensemble learning methods.

No. 12KJB520003 and China Postdoctoral Science Foundation under grant No. 2013M540404.

REFERENCES

[1] N.V. Chawla, N. Japkowicz, and A. Kolcz, "Editorial: Special Issue on Learning from Imbalanced Data Sets," *ACM SIGKDD Exploration Newsletter*, vol. 6, no. 1, pp. 1-6, 2004.

[2] N. Japkowicz, "Learning from Imbalanced Data Sets," *Proc. Am. Assoc. for Artificial Intelligence (AAAI) Workshop*, 2000.

[3] N.V. Chawla, N. Japkowicz, and A. Kolcz, "Workshop Learning from Imbalanced Data Sets II," *Proc. Int'l Conf. Machine Learning*, 2003.

[4] N.V. Chawla, K.W. Bowyer, and L.O. Hall, "SMOTE: Synthetic Minority Over-Sampling Technique," *J. Artificial Intelligence Research*, vol. 16, no. 1, pp. 321-357, 2002.

[5] H. Han, W.Y. Wang, and B.H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," *Proc. Int'l Conf. Intelligent Computing (ICIC'05)*, pp. 878-887, 2005.

[6] P. Yang, L. Xu, and B. Zhou, "A Particle Swarm Based Hybrid System for Imbalanced Medical Data Sampling," *BMC Genomics*, vol. 10, no. S3, article S34, 2009.

[7] S.J. Yen and Y.S. Lee, "Cluster-Based Under-Sampling Approaches for Imbalanced Data Distributions," *Expert Systems with Applications: An Int'l J.*, vol. 36, no. 3, pp. 5718-5727, 2009.

[8] Z.H. Zhou and X.Y. Liu, "Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem," *IEEE Trans. Knowledge Data Eng.*, vol. 18, no. 1, pp. 63-77, Jan. 2006.

[9] C. Elkan, "The Foundations of Cost-Sensitive Learning," *Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI'01)*, pp. 973-978, 2001.

[10] Y. Sun, M.S. Kamel, and A.K.C. Wong, "Cost-Sensitive Boosting for Classification of Imbalanced Data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358-3378, 2007.

[11] N.V. Chawla, A. Lazarevic, and L.O. Hall, "SMOTEBoost: Improving Prediction of the Minority Class in Boosting," *Proc. European Conf. Principles of Data Mining and Knowledge Discovery (PKDD'03)*, pp. 107-119, 2003.

[12] X.Y. Liu, J. Wu, and Z.H. Zhou, "Exploratory Undersampling for Class-Imbalance Learning," *IEEE Trans. Systems Man and Cybernetics -Part B*, vol. 39, no. 2, pp. 539-550, Apr. 2009.

[13] D. Tao, X. Tang, and X. Li, "Asymmetric Bagging and Random Subspace for Support Vector Machines-Based Relevance Feedback in Image Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 7, pp. 1088-1099, July 2006.

[14] G.Z. Li, H.H. Meng, and W.C. Lu, "Asymmetric Bagging and Feature Selection for Activities Prediction of Drug Molecules," *BMC Bioinformatics*, vol. 9, no. S6, article S7, 2008.

[15] H.J. Shin, D.H. Eom, and S.S. Kim, "One-Class Support Vector Machines-An Application in Machine Fault Detection and Classification," *Computers & Industrial Eng.*, vol. 48, no. 2, pp. 395-408, 2005.

[16] K.K. Seo, "An Application of One-Class Support Vector Machines in Content-Based Image Retrieval," *ExpertSystems with Applications*, vol. 33, no. 2, pp. 491-498, 2007.

[17] X. Hong, S. Chen, and C.J. Harris, "A Kernel-Based Two-Class Classifier for Imbalanced Data Sets," *IEEE Trans. Neural Networks*, vol. 18, no. 1, pp. 28-41, Jan. 2007.

[18] G. Wu and E.Y. Chang, "KBA: Kernel Boundary Alignment Considering Imbalanced Data Distribution," *IEEE Trans. Knowledge and Data Eng.*, vol. 17, no. 6, pp. 786-795, June 2005.

[19] S. Ertekin, J. Huang, and C.L. Giles, "Active Learning for Class Imbalance Problem," *Proc. Int'l SIGIR Conf. Research and Development in Information Retrieval*, pp. 823-824, 2007.

[20] S.Y. Oh, M.S. Lee, and B.T. Zhang, "Ensemble Learning with Active Example Selection for Imbalanced Biomedical Data Classification," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 8, no. 2, pp. 316-325, Mar./Apr. 2011.

[21] R. Blagus and L. Lusa, "Class Prediction for High-Dimensional Class-Imbalanced Data," *BMC Bioinformatics*, vol. 11, article 523, 2010.

[22] W.J. Lin and J.J. Chen, "Class-Imbalanced Classifiers for High-dimensional Data," *Brief Bioinformatics*, vol. 14, no. 1, pp. 13-26, 2013.

[23] M. Wasikowski and X.W. Chen, "Combating the Small Sample Class Imbalance Problem Using Feature Selection," *IEEE Trans. Knowledge and Data Eng.*, vol. 22, no. 10, pp. 1388-1400, Oct. 2010.
[24] T.M. Khoshgoftaar, J.V. Hulse, and A. Napolitano, "Comparing Boosting and Bagging Techniques with Noisy and Imbalanced Data," *IEEE Trans. Systems, Man, and Cybernetics-Part B*, vol. 41, no. 3, pp. 552-568, Mar. 2011.
[25] A.H.M. Kamal, X. Zhu, and R. Narayanan, "Gene Selection for Microarray Expression Data with Imbalanced Sample Distributions," *Proc. Int'l Joint Conf. Bioinformatics, System Biology and Intelligent computing*, pp. 3-9, 2009.
[26] H.L. Yu, J. Ni, and J. Zhao, "ACOSampling: An Ant Colony Optimization-Based Undersampling Method for Classifying Imbalanced DNA Microarray Data," *Neurocomputing*, vol. 101, no. 2, pp. 309-318, 2013.
[27] P. Yang, Z. Zhang, B.B. Zhou, and A.Y. Zomaya, "A Clustering Based Hybrid System for Biomarker Selection and Sample Classification of Mass Spectrometry Data," *Neurocomputing*, vol. 73, no. 13/15, pp. 2317-2331, 2010.
[28] I. Levner, "Feature Selection and Nearest Centroid Classification for Protein Mass Spectrometry," *BMC Bioinformatics*, vol. 6, article 68, 2005.
[29] N.G. Pedrajas, J.P. Rodríguez, and M.G. Pedrajas, "Class Imbalance Methods for Translation Initiation Site Recognition in DNA Sequences," *Knowledge-Based Systems*, vol. 25, no. 1, pp. 22-34, 2012.
[30] R. Batuwita and V. Palade, "MicroPred: Effective Classification of Pre-miRNAs for Human miRNA Gene Prediction," *Bioinformatics*, vol. 25, no. 8, pp. 989-995, 2009.
[31] V. Vapnik, *Statistical Learning Theory*. Wiley Publishers, 1998.
[32] X.M. Zhao, X. Lin, L. Chen, and K. Aihara, "Protein Classification with Imbalanced Data," *Proteins: Structure, Function and Bioinformatics*, vol. 70, pp. 1125-1132, 2008.
[33] L.A. Kurgan, K.J. Cios, R. Tadeusiewicz, M. Ogiela, and L. Goodenday, "Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis," *Artificial Intelligence in Medicine*, vol. 23, no. 2, pp. 149-169, 2001.
[34] H. Liu, H. Han, J. Li, and L. Wong, "An In-Silico Method for Prediction of Polyadenylation Signals in Human Sequences," *Proc. 14th Int'l Conf. Genome Informatics*, pp. 84-93, 2003.
[35] R.J. Dobson, P.B. Munroe, M.J. Caulfield, and M.A.S. Saqi, "Predicting Deleterious nsSNPs: An Analysis of Sequence and Structural Attributes," *BMC Bioinformatics*, vol. 7, pp. 217-225, 2006.
[36] A. Ozcift, "Random Forests Ensemble Classifier Trained with Data Resampling Strategy to Improve Cardiac Arrhythmia Diagnosis," *Computers in Biology and Medicine*, vol. 41, pp. 265-271, 2011.
[37] L. Breiman, "Bagging Predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.
[38] T. Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832-844, Aug. 1998.
[39] A. Krogh and J. Vedelsby, "Neural Network Ensembles, Cross Validation, and Active Learning," *Advances in Neural Information Processing Systems*, vol. 7, pp. 231-238, 1995.
[40] Y.H. Wang, F.S. Makedon, J.C. Ford, and J. Pearlman, "HykGene: A Hybrid Approach for Selecting Feature Genes for Phenotype Classification Using Microarray Gene Expression Data," *Bioinformatics*, vol. 21, no. 8, pp. 1530-1537, 2005.
[41] T.R. Golub, D.K. Slonim, and P. Tamayo, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, no. 5439, pp. 531-537, 1999.
[42] S.B. Cho and H.H. Won, "Cancer Classification Using Ensemble of Neural Networks with Multiple Significant Gene Subsets," *Applied Intelligence*, vol. 26, no. 3, pp. 243-250, 2007.
[43] U. Alon, N. Barkai, and D.A. Notterman, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Array," *Proc. Nat'l Academy of Science USA.*, vol. 96, no. 12, pp. 6745-6750, 1999.
[44] D.A. Wigle, I. Jurisica, and N. Radulovich, "Molecilar Profiling of Non-Small Cell Lung Cancer and Correlation with Diseasefree Suvival," *Cancer Research*, vol. 62, no. 11, pp. 3005-3008, 2002.
[45] E.F. Petricoin, A.M. Ardekani, and B.A. Hitt, "Use of Proteomic Patterns in Serum to Identify Ovarian Cancer," *The Lancet*, vol. 359, no. 9306, pp. 572-577, 2002.
[46] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.
[47] S. Keerthi and C.J. Lin, "Asymptotic Behaviours of Support Vector Machines with Gaussian Kernel," *Neural Computing*, vol. 15, no. 7, pp. 1667-1689, 2003.
[48] E.K. Tang, P.N. Suganthan, and X. Yao, "An Analysis of Diversity Measures," *Machine Learning*, vol. 65, pp. 247-271, 2006.

Hualong Yu received the BS degree from Heilongjiang University, Harbin, China, in 2005, and the MS and PhD degrees from the College of Computer Science and Technology, Harbin Engineering University, Harbin, China, in 2008 and 2010, respectively. Since 2010, he has been one lecturer and master supervisor in Jiangsu University of Science and Technology, Zhenjiang, China. He is author or co-author for more than 30 research papers and three books. He is also the program committee member for ICICSE2012, ICICSE2013, and the reviewer for more than 15 professional journals. His research interests mainly include machine learning and Bioinformatics.

Jun Ni received the BS degree from Harbin Engineering University, Harbin, China, the MS degree from Shanghai Jiaotong University, Shanghai, China, and the PhD degree from the University of Iowa in 1981, 1984 and 1991, respectively. He is currently an associate professor and director of Medical Imaging HPC and Informatics Lab, Department of Radiology, Carver College of Medicine, the University of Iowa, Iowa City. He is also one visiting professor in Harbin Engineering University, Shanghai University and Nanjing University of Science and Technology, China, since 2006, 2009 and 2012, respectively. He edited or co-edited 34 books or proceedings and authored or co-authored 120 peer-reviewed journal and conference papers. In addition, he is editor-in-chief of *International Journal of Computational Medicine and Healthcare*, associate editor of *IEEE Systems Journal* and editorial board member for 15 other professional journals. Since 2003, he has also been General/Program chairs for more than 50 International conferences. Currently, his research interests include distributed computation, parallel computing, medical imaging informatics, computational biology, and Bioinformatics, etc. He is also a member of the IEEE.