Available online at www.ijpe-online.com

International Journal of Performability Engineering

Vol. 13, No. 5, September 2017, pp. 742-753
DOI: 10.23940/ijpe.17.05.p17.742753

A Novel Information Theory-Based Ensemble Feature Selection Framework for High-Dimensional Microarray Data

Jie Cai[a], Jiawei Luo[a,*], Cheng Liang[b], ShengYang[a]

[a]*College of Computer Science and Electronic Engineering, Hunan University, Changsha, 410082, Hunan, China*
[b]*School of Information Science and Engineering, Shandong Normal University, Jinan, 250358, Shangdong, China*

Abstract

Ensemble feature selection is one of the ensemble learning methods, where each classifier is trained or built by feature selection result. Ensemble feature selection is an effective way for dealing with high dimension and small sample data, such as microarray data. However, ensemble feature selection should achieve more accurate and stable classification performance. In this paper, we present a novel diversity measure based on information theory called Sum of Minimal Information Distance (SMID), which maximizes the relevance between feature subsets and class label as well as the diversity between feature subsets. Moreover, a novel ensemble feature selection framework satisfying this criterion is proposed. In this framework, features that have more mutual information with class label and more diversity between each other are retained. Different feature subsets are used to train base classifiers after being obtained by incremental search method, and then these classifiers are aggregated into a consensus classifier by majority voting. Comparing with three representative feature selection methods and five ensemble learning methods on ten microarray datasets, the experiment results show that the proposed method achieves better performance than the other methods in terms of the classification accuracy.

Keywords: Classification; Feature selection; Ensemble learning; Diversity; Mutual information

(Submitted on March 8, 2017; Revised on July 1, 2017; Accepted on August 27, 2017)

1. Introduction

Microarray data often presents high dimensionality, small samples and imbalance. A large number of ensemble learning methods have been applied to microarray data classification [1,16,19,21]. Ensemble classification is the process of combining multiple models and aggregating their results into a single decision [9,29]. It confers a number of benefits, including more robust and flexible classification models as well as improved classification performances. In general, the performance of ensemble learning depends on the performance of the base classifier and diversity between the base classifiers [14]. The greater the diversity between the base classifiers, the higher the generalization ability of the ensemble learning system. Consequently, the improvement of its performance effect would also be guaranteed. On the contrary, if each base classifier adopts the same structure and the degree of diversity between them is low, the performance effect of ensemble learning system would be greatly reduced. Therefore, an ensemble learning algorithm that provides good performance and strong generalization ability is usually designed by improving the diversity between the base classifiers. As a whole, the diversity between the base classifiers is mainly embodied in the following three aspects: different training samples, different base classifiers and different feature spaces.

Ensemble feature selection is the ensemble of base classifiers trained by different feature spaces, which are generated by feature selection algorithms. It can improve the stability of feature selection and obtain the diverse training data to improve the performance of the ensemble classifiers [15,17,23]. For Microarray data, it generally has small size of samples, so ensemble method using different training samples might degrade the classification performance of base classifiers, while ensemble feature selection can manipulate multiple gene sets simultaneously and maintain the performance of base classifiers.

* Corresponding author.
E-mail address: luojiawei@hnu.edu.cn.

In recent years, the ensemble of different feature spaces and ensemble feature selection has become an important focus of research. Ho [11] was enlightened by the theory of Stochastic Discrimination (SD) and proposed the Random Subspace Method (RSM) for constructing decision forests. Since then, this type of method has attracted great interest. Random forest algorithm is considered as the representative method using this idea [6]. Ahn et al. [3] adopted the random method to partition the feature space into mutually exclusive subspaces for generating base classifiers. They introduced an ensemble-based approach for classification called CERP, which was designed specifically for high-dimensional gene expression datasets. Bock [8] introduced Generalized Additive Models (GAMs) as base classifiers for binary ensemble classification using RSM and/or Bagging. Three alternative ensemble strategies using GAMs as base classifiers were proposed: GAMbag, GAMrsm and GAMens. The proposed GAMs perform well in a binary classification, not in multi-class problems. Liu [16] proposed a new ensemble gene selection method based on information theory to obtain multiple gene subsets by the same selection technique with different starting points in its search procedure. The ensemble method yields high accuracy, but its computational cost is relatively higher than other filters. Additionally, the quantity of base classifiers is different for different datasets. Zhang [28] proposed a novel transformation based method to increase the diversity of each tree in the forests to improve the overall accuracy. The method improves the performance of the Random Forests in most cases, but involves a large number of different types of decision trees.

Ensemble feature selection aims to reduce the influence of high dimension on learning algorithm and produce the ensemble with diverse base classifiers simultaneously, and then build effective ensemble learning algorithm that is suitable for classification problem. As ensemble feature selection is essentially the combination of classifier ensemble and feature selection, it could achieve excellent performance when dealing with high-dimensional microarray data [4,16]. However, the performance of these ensemble methods is not always stable. This instability is caused by the fact that the feature subset is randomly divided and the diversity between the selected feature subsets is not guaranteed. In this paper, we propose the information diversity measure between selected feature subsets as the diversity of base classifiers. The sum of minimal information distance measure (SMID) is presented as an alternative of the information diversity measure between selected feature subsets since it is hard to calculate. We then use SMID to design a novel ensemble feature selection framework to achieve the diversity between feature subsets and the diversity between base classifiers that are trained by these feature subsets. After the base classifiers finished training, we choose the basic classifiers with high accuracy to build the ensemble classifier according to the majority voting rules. To verify the performance of our method, we compare the proposed ensemble framework with three representative feature selection methods as well as five ensemble classification methods on ten high-dimensional microarray datasets. The results demonstrate that the proposed framework could obtain better performance over the other methods in terms of classification accuracy in most cases.

2. Information theory analysis of ensemble feature selection

2.1. Maximizing relevance and diversity

For a given dataset $D=(X,C)$, with a feature set $X=\{x_1,x_2,...,x_n\}$ and class label C, the ensemble feature selection is to obtain K feature subsets $\Xi=\{X_1, X_2, \ldots, X_K\}$. Each feature subset $X_i \subset \Xi$ is first used to train a base classifier and then integrated into an overall outcome (ensemble model) by some combination strategies, such as majority voting, weighting voting and function method [5,27]. The common framework is shown in Figure 1.

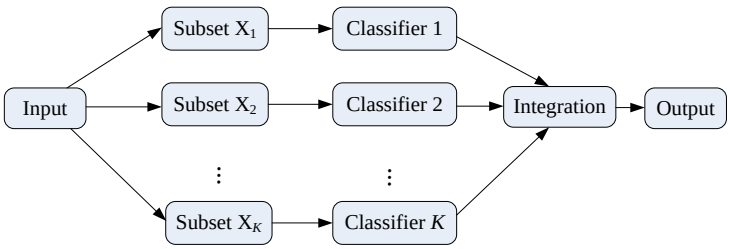

Figure 1. The common framework of the ensemble feature selection

The ensemble learning emphasizes good classification ability of the base classifier and great diversity between them. Krogh[13] pointed out that error function of ensemble can be decomposed as formula (1):

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

In formula (1), E is the ensemble error, $\overline{E}$ is the average error of base classifiers and $\overline{A}$ is the ensemble diversity. Formula (1) shows that when the diversity is enlarged for a fixed average error, the generalization error decreases, in other words, the system generalization ability increases. However, the average error is correlated with the diversity: the average error will increase/decrease as the diversity increases/decreases. Therefore, how to find the balance state between the average error (or precision) and the diversity is crucial to improve the generalization ability.

From the point of view of information theory, the aim of ensemble feature selection is to first find a set of feature subsets with great diversity information between each other while each subset has as much information as possible with the target class, and finally to obtain a good ensemble classifier. The mutual information between the feature subset and the target class, i.e. $I(X_i;C)$, can effectively measure the relevance between X_i and C. The higher $I(X_i;C)$ and the relevance are, the smaller the classification error probability of the classifier is. The idea of maximizing the mutual information between the feature subset and the target class, i.e. $\max I(X_i;C)$, has achieved great success in mutual information-based feature selection algorithms for classification, such as MRMR[18], JMI[26], CMIM[10] and etc. Moreover, the diversity between the feature subsets is maximized as great as possible, i.e., $\max D(X_i, X_j)$, where $D(.)$ is the information diversity measure. Taken together, the ensemble feature selection can be expressed by information theory as follows:

$$\min(E) \equiv \max[I(X_i; C)] \text{ AND } \max[D(X_i, X_j)] \tag{2}$$

2.2. Sum of minimal information distance

Vinh et al.[24] theoretically analyzed the variant of information $d(x_i, x_j)$ of two random variables. $d(x_i, x_j)$ is defined as $H(x_i \mid x_j) + H(x_j \mid x_i)$ or $H(x_i, x_j) - I(x_i; x_j)$, where $H(x_i \mid x_j)$ and $H(x_j \mid x_i)$ are conditional entropies, and $H(x_i, x_j)$ is the joint entropy. It is proved to be a standard information theory-based distance measure or diversity measure, satisfying symmetry, non-negative and triangle inequality, and on which a feature space can be built. To expand the measure to the feature subset, given two feature subsets X_i and X_j, the information diversity measure can be expressed as follows:

$$D(X_i, X_j) = H(X_i \mid X_j) + H(X_j \mid X_i) = H(X_i, X_j) - I(X_i; X_j) \tag{3}$$

In general, it is intractable to calculate $H(X_i \mid X_j)$ or $H(X_j \mid X_i)$ as same as joint entropy $H(X_i, X_j)$ and joint mutual information $I(X_i; X_j)$. In many information theory-based algorithm designs, the sum or the average of individual measure is often approximately adopted as the measure of feature subset. Here, we define a new information diversity measure between feature subsets called Sum of Minimal Information Distance (SMID). Given two feature subsets $A = \{a_1, a_2, ..., a_p\}$ and $B = \{b_1, b_2, ..., b_q\}$, $A, B \subset \Xi$, SMID is referred to as the sum of minimal information distance between each feature from one subset and all features from another subset. Its definition can be formulated as follows:

$$SMID(A, B) = \sum_{a_i \in A} \min_{b_j \in B} d(a_i, b_j) + \sum_{b_j \in B} \min_{a_i \in A} d(b_j, a_i) \tag{4}$$

For instance, let $A=\{a_1, a_2, a_3, a_4\}$ and $B=\{b_1, b_2, b_3, b_4\}$, the minimum information distance between A and B is given in Figure 2. For a_3, we get $\min_{b_j \in B} d(a_3, b_j) = d(a_3, b_4)$. For a_2 and b_3, we get $\min_{b_j \in B} d(a_2, b_j) = \min_{a_i \in A} d(b_3, a_i) = d(a_2, b_3)$. Taken together, we have $SMID(A,B)=d(a_1,b_3)+d(a_2,b_3)+d(a_3,b_4)+d(a_4,b_4)+d(b_1,a_1)+d(b_2,a_1)+d(b_3,a_2)+d(b_4,a_2)$.

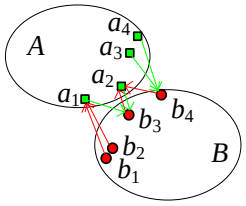

Figure 2. SIMD between A and B

According to the definition of SMID, *SMID*(*A*, *B*) has the following properties:

- It is always non-negative: $SMID(A, B)\geq 0$.
- If $A=B$, then $SMID(A, B)=0$.
  Proof.
  Since $A=B$, we have
  $\min_{b_j\in B} d(a_i,b_j)=d(a_i,a_i)=0$ and $\min_{a_i\in A} d(a_i,b_j)=d(b_j,b_j)=0$
  Thus, $SMID(A, B)=0$
- It is commutative: $SMID(A, B)= SMID(B, A)$.
- $SMID(A, B)\geq D(A, B)$
  Proof.
  Let $A=\{a_1,a_2,...,a_p\}$ and $B=\{b_1,b_2,...,b_q\}$, according to the definition and the properties of conditional entropy, we have

$$H(A\mid B)=H(\{a_1,a_2,...,a_p\}\mid B)\leq \sum_{a_i\in A} H(a_i\mid B)$$

$$H(B\mid A)=H(\{b_1,b_2,...,b_q\}\mid A)\leq \sum_{b_j\in B} H(b_j\mid A)$$

$$H(a_i\mid B)=H(a_i\mid\{b_1,b_2,...,b_q\})\leq \min_{b_j\in B} H(a_i\mid b_j)$$

$$H(b_j\mid A)=H(b_j\mid\{a_1,a_2,...,a_p\})\leq \min_{a_i\in A} H(b_j\mid a_i)$$

Thus,

$$\begin{aligned}
&D(A,B)=H(A\mid B)+H(B\mid A)\\
&\leq \sum_{a_i\in A} H(a_i\mid B)+\sum_{b_j\in B} H(b_j\mid A)\\
&\leq \sum_{a_i\in A}\min_{b_j\in B} H(a_i\mid b_j)+\sum_{b_j\in B}\min_{a_i\in A} H(b_j\mid a_i)\\
&\leq \sum_{a_i\in A}\min_{b_{j'}\in B}(H(a_i\mid b_{j'})+H(b_{j'}\mid a_i))+\sum_{b_j\in B}\min_{a_{i'}\in A}(H(b_j\mid a_{i'})+H(a_{i'}\mid b_j))\\
&=\sum_{a_i\in A}\min_{b_{j'}\in B} d(a_i,b_{j'})+\sum_{b_j\in B}\min_{a_{i'}\in A} d(b_j,a_{i'})\\
&=SMID(A,B)
\end{aligned}$$

Unlike traditional diversity measures using the classification results of base classifiers to calculate the diversity[14], SMID uses these features to calculate the diversity between feature subsets, which denotes the diversity between generated classifiers. Since SMID can be calculated directly, we take it as an alternative of $D(A, B)$. Finally, the approximation of formula (2) is given as follows:

$$\min(E)\equiv\max[I(X_i; C)] \text{ AND } \max[SMID(X_i, X_j)] \tag{5}$$

3. Ensemble feature selection framework based on SMID

As seen from formula (5), ensemble feature selection aims to find the feature subsets with greater relevance to class label and more diversity between the subsets. A good feature subset is normally formed by some features with greater relevance to class label, i.e. they have higher mutual information. According to the definition of SMID, increasing the information distance between the features from different subsets can increase SMID. Therefore, a better ensemble feature selection can be obtained by grouping properly the features with higher mutual information to class label and greater information distance between each other.

A new Ensemble feature selection Framework based on SMID (EFSMID) is proposed. First, the features with less classification ability and smaller information distance to other features are removed for Increasing the Information Distance (IID) between features; next, the feature space is divided into several groups according to the feature selection criterion J_*, and then several feature subsets are obtained to train the base classifiers; finally, these classifiers are integrated to get an overall outcome by majority voting. The steps of EFSMID can be summarized as: (1) IID; (2) grouping; (3) ensembling. Details are given as follows:

Algorithm: EFSMID

Input: ***D*** - dataset with original feature set X and class ***C***
α - IID coefficient
K - the object number of feature subsets for grouping
$\hat{k}$ - the object number of features in each selected feature subset
Output: ensemble classifier
Steps:

Step1 IID

1. Calculate the information distance d_{ij} between $x_i \in X$ and $x_j \in X$.
2. Calculate the average information distance d_{avg}=sum(d_{ij})/(n-1)^2.
3. **for** each feature $x_i \in X$ **do**
   Calculate $I(x_i;C)$
   end
4. Sort them in the descending order according to $I(x_i;C)$.
5. **for** each feature $x_i \in X$ **do**
   for each feature $x_j \in X$, $j > i$ **do**
   if $d_{ij} < \alpha d_{avg}$ **then**
   $X \leftarrow X \setminus \{x_j\}$
   end
   end
   end

Step2 grouping

1. Initialize K feature subsets, $X_i \leftarrow \phi,\ i = 1,2,...,K$
2. **for** k=1 to $\hat{k}$ **do**
   for i=1 to K **do**
   if $x_j \in X$ satisfying the incremental search criterion J_* **then**
   $X_i \leftarrow X_i \cup \{x_j\}$
   $X \leftarrow X \setminus \{x_j\}$
   end
   end
   end

Step3 ensembling

1. Train the base classifier BC_i by the obtained feature subsets X_i, i=1, … , K.
2. Calculate the classification accuracy Acc_i of BC_i and the average accuracy Acc_{avg}.
3. Apply majority voting method on the base classifiers satisfying $Acc_i >= Acc_{avg}$, output the ensemble classifier.

IID removes the features that are close to other features with greater $I(x_i;C)$. The IID coefficient α decides the minimal distance between the features. The higher the α is, the greater the distance as well as the SMID between the feature subsets will be. However, a higher α will also lead to more candidate features being removed. Specifically, an excessive IID coefficient may decrease the value of $I(X_i;C)$, and the performance of the base classifier might also be decreased after the features are grouped. In contrast, a moderate IID coefficient can not only increase the diversity between feature subsets, but also enhance the combination ability between features as well as the mutual information between feature subsets and class label C. Generally, as the information diversity between the features increases, the information redundancy will decrease and the classifier is more likely to get higher classification ability. To satisfy both $\max[I(X_i;C)]$ and $\max[SMID(X_i,X_j)]$ simultaneously, we adjust the IID coefficient to search for the best ensemble classifier of EFSMID, which gives a well trade-off between the accuracy and diversity of base classifiers.

In the grouping step, features will be assigned to different feature subsets according to the incremental search criterion J_* sequentially till each feature subset includes $\hat{k}$ features or there is no feature to be assigned. There are $\hat{k}$ rounds of assignment, and K features are assigned to K feature subsets in each round, respectively. Then each feature subset is used to

train a base classifier. Three classic incremental search criterions based on information theory are used to group the features, which are MRMR, JMI and CMIM.

In ensembling, the base classifiers with better performance are selected into the ensemble classifier using majority voting principle. Although IID can guarantee the diversity of base classifiers, some poor base classifiers can still be generated especially when the IID coefficient is large. Therefore, EFSMID only retains the better base classifiers that reduce the value of $\overline{E}$ in formula (1). Generally, there are $K/2$ base classifiers for voting since EFSMID uses $Acc_i \geq Acc_{avg}$ to select the base classifier.

The time complexity of EFSMID depends on the time complexity of IID, grouping and ensembling of the base classifiers. For IID, it needs to first calculate the distance between the features, sort the features according to $I(x_i;C)$ and then remove the features according to α. Suppose the dataset includes m samples and n features, the time complexity of calculating the information distance is $O(mn^2)$, while the sorting is $O(n^2)$ and the removing is $O(n^2)$. Therefore, the time complexity of IID is mainly decided by the calculation of the information distance, i.e., $O_1(mn^2)$. The complexity of grouping depends on the complexity of feature selection method. As aforementioned, the ensemble framework generates K target subsets and each target subset includes $\hat{k}$ features. Since the value of $I(x_i;C)$ and $d(x_i,x_j)$ have been computed, the time complexity of grouping is $O_2(K\hat{k}n^2)$. The complexity of the base classifier ensembling is mainly from the classification learning algorithm. If the complexity of the classification learning algorithm is O_3, then the overall complexity is KO_3. In summary, EFSMID has a polynomial time complexity as long as the time complexity of the feature selection method and the classification learning algorithm are polynomial.

4. Experiments

To test the performance of the proposed ensemble framework, ten cancer microarray datasets are selected to validate the classification accuracy. Three of the datasets (Breast, CNS and Colon Tumor) can be accessed from Kent Ridge Bio-medical Dataset website (http://datam.i2r.a-star.edu.sg/datasets/krbd/). NCI datasets can be downloaded from (http://ligarto.org/rdiaz/Papers/rfVS/randomForestVarSel.html). The rest can be downloaded from GEMS website (http://www.gems-system.org/). Details of the ten datasets are summarized in the following table. Note that these ten datasets differ greatly in the sample sizes (ranging from 50 to 174) as well as the number of features (ranging from 2,000 to 126,001). Furthermore, the number of classes varies from 2 to 11, where '2' indicates a binary classification and other numbers indicate multi-class problems. These datasets can provide a comprehensive test for EFSMID under different conditions. For continuous and mixed datasets, MDL discretization method [12] is used to transform continuous features into discrete ones.

Table 1. Description of datasets

	Dataset	Samples	Features	Classes
1	9 Tumors	60	5727	9
2	BreastCancer	97	24481	2
3	Prostate Cancer	136	126001	2
4	BrainTumor	90	5921	5
5	11 Tumors	174	12533	11
6	NCI	61	5244	8
7	ColonTumor	62	2000	2
8	CNS	60	7129	2
9	Gliomas	50	12625	2
10	SRBCT	83	2308	4

Since EFSMID is based on information theory, we carry out experimental comparisons with four representative feature selection methods, i.e. MRMR, JMI, CMIM and MRMD [30]. The first three methods are based on information theory. MRMD is a feature ranking method based on the distance function. EFSMID is combined with each of the three methods, and three specific ensemble feature selection methods EFSMID+MRMR, EFSMID+JMI and EFSMID+CMIM are obtained. All the EFSMID methods are implemented in Matlab and Weka. In EFSMID, we set K=10 and $\hat{k}=10$, which means that there are 10 base classifiers for selection in ensembling and each base classifier is trained with 10 selected features. In order to check the ensemble ability of EFSMID, the number of selected features in the above three feature selection methods is set as 10. Besides, other kinds of ensemble classification methods including Adaboost M1, Bagging, Randomforest[6], RandomSubSpace[11] and RotationForest[22] are also compared with the proposed methods. Five of the ensemble methods to be compared have already been integrated into Weka, so we can directly use them in Weka to obtain their classification accuracy. Four widely used Classifiers-Naive Bayes Classifier (NBC) [25], Support Vector Machines (SVM) [7], Nearest-Neighbour (IB1)[2] and Decision Tree C4.5 (J48)[20] are employed as the standard classifiers to generate classification

accuracy. Classification accuracy is the most widely used metric in classification and it is defined as $Acc = |samples\ correctly\ classified| / |all\ samples|$. All classification accuracies are obtained by ten times 10-fold cross-validation. Similarity, the F-measure (F_1) is also the important metric and it is defined as F_1=(2*precision*recall)/(precision+recall).

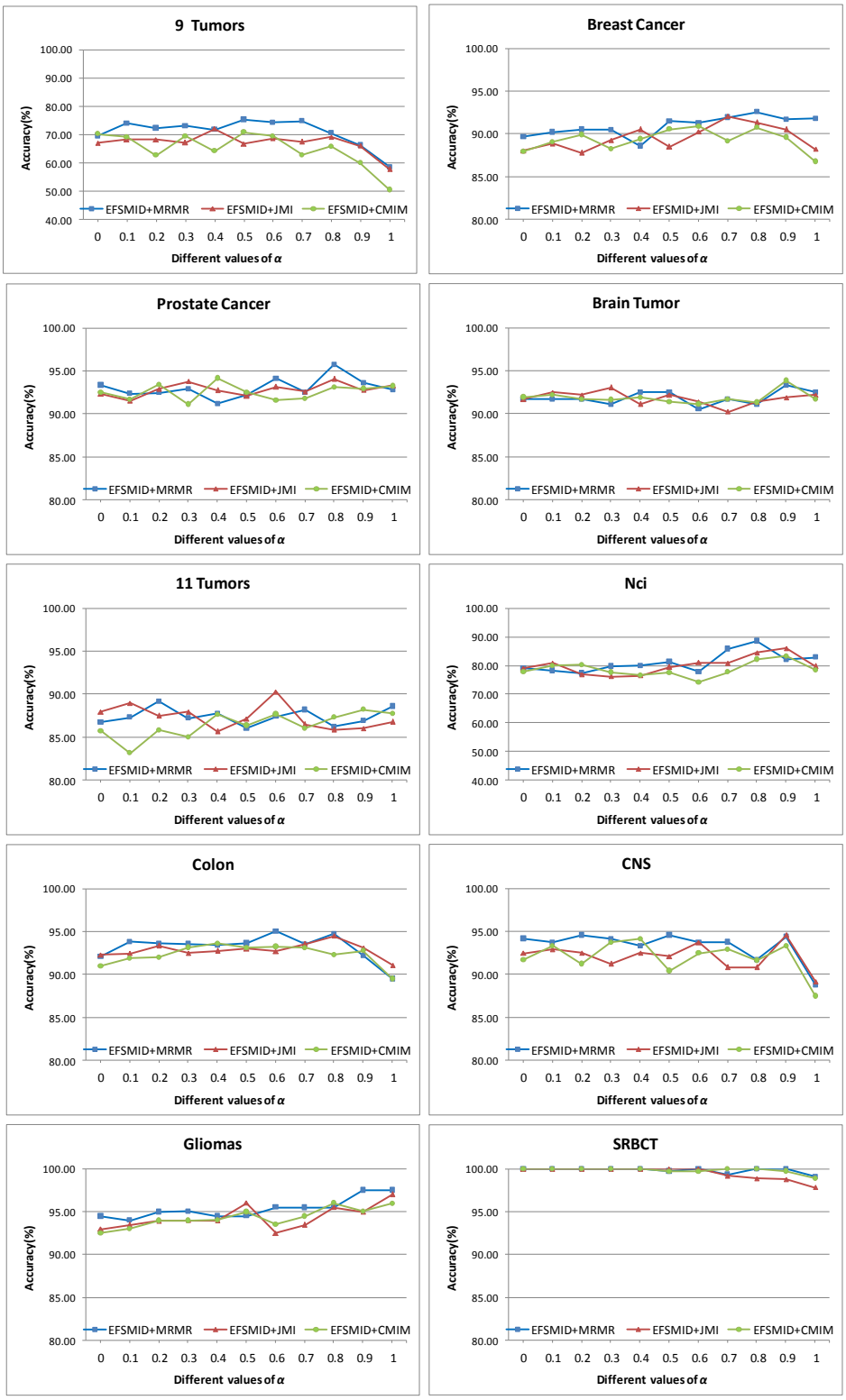

Figure 3. The average accuracy varies with different values of α.

Figure 3 shows the average classification accuracy of EFSMID under the four kinds of classification learning methods with different values of α. From Figure 3, we can see that the best ensemble classifier presents in $\alpha \in [0.1, 0.9]$ in most cases. That means increasing feature information distance could obtain base classifiers with more diversity and better classification performance. When α=0, no features are removed from the original feature set and too many redundancy features will be used for grouping, resulting in less diversity between feature subsets and also more feature-feature redundancy in the incremental search processing. However, too many features that have high mutual information with the class label are deleted if α=1. Although the diversity between the base classifiers is increased in this case, the classification accuracy of base classifier is decreased significantly. In Figure 3, the ensemble classifiers show the best performance when α is in the range of [0.4, 0.6] on 9 Tumors, 11 Tumors, ColonTumor and CNS, respectively. However, for BreastCancer, ProstateCancer, BrainTumor, NCI and Gliomas, the appropriate α is [0.7, 0.9], which is greater than that on other datasets. The classification accuracy of EFSMID fluctuates in terms of α, so the IID coefficient can be adjusted to search the best trade-off between the diversity and the performance of base classifiers.

Table 2 shows the best results of the ensembles obtained by different classification learning methods over the datasets. From Table 2, it can be seen that almost all the best classification accuracies are obtained by EFSMID. The results demonstrate the adequacy of the proposed ensemble framework, since they match or improve upon the results achieved by the filters alone. EFSMID wins all ten datasets when C4.5 is used as the classification learning method. For most of the multiple classification datasets, such as 9 Tumors, Brain Tumors, 11 Tumors and NCI, EFSMID exhibits an obvious improvement over the results achieved by the filters. For SRBCT dataset, EFSMID can almost ignore the diversity between the base classifiers since they have very excellent classification performance. We also compare F_1 for each method in

Table 3. Here, it can be seen that almost all the best F_1 are obtained by EFSMID. This indicates that EFSMID has superior classification capability over the other feature methods.

Table 2. Test classification accuracy for microarray datasets. The number in parenthesis is the value of α on best accuracy gained by EFSMID. Best accuracy for each dataset is highlighted in bold face.

Dataset	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
NBC Ensembles										
EFSMID+MRMR	**86.10**	93.92	**92.75**	95.50	**93.28**	**95.55**	**96.67**	**98.30**	98.10	**100**
	(0.4)	(1.0)	(0.8)	(0.9)	(0.2)	(0.8)	(0.8)	(0.5)	(1.0)	(0~0.9)
EFSMID+JMI	83.15	**94.10**	91.24	**96.68**	92.42	95.12	95.43	95.08	98.08	**100**
	(0.6)	(0.7)	(0.8)	(0.3)	(0.6)	(0.9)	(0.8)	(0.9)	(1.0)	(0~0.8)
EFSMID+CMIM	85.05	91.80	91.77	96.65	93.13	93.30	95.20	**98.30**	98.10	**100**
	(0.5)	(0.5)	(0.4)	(0.9)	(0.6)	(0.8)	(0.5)	(0.4)	(0.9)	(0~0.9)
Filters										
MRMR	60.33	87.62	91.15	90.10	85.10	77.15	95.08	95.50	**98.20**	98.78
JMI	70.12	84.55	91.92	92.25	83.92	70.40	93.07	98.12	97.81	98.74
CMIM	70.10	87.68	77.98	91.17	86.27	78.60	95.04	98.16	95.43	98.72
MRMD	60.20	84.54	86.76	84.44	82.75	75.41	87.09	95.60	94.56	97.59
SVM Ensembles										
EFSMID+MRMR	**65.05**	**95.08**	96.33	**92.25**	88.15	**80.74**	**95.43**	95.10	**98.10**	**100**
	(0.5)	(1.0)	(0.8)	(0.9)	(0.2)	(0.8)	(0.6)	(0.6)	(1.0)	(0~0.6)
EFSMID+JMI	60.03	92.80	95.77	91.10	**90.70**	78.31	95.06	95.09	96.07	**100**
	(0.4)	(0.8)	(0.9)	(0.3)	(0.6)	(0.8)	(0.8)	(0.9)	(1.0)	(0~0.5)
EFSMID+CMIM	58.32	92.82	**97.08**	92.22	89.15	77.10	93.51	96.60	**98.10**	**100**
	(0.8)	(0.6)	(0.4)	(0.5)	(0.9)	(0.9)	(0.4)	(0.4)	(1.0)	(0~0.6)
Filters										
MRMR	56.62	82.46	92.62	88.80	87.32	75.42	93.51	95.30	95.80	98.79
JMI	60.67	83.52	90.07	90.04	85.76	75.42	95.18	**98.52**	95.80	97.59
CMIM	56.62	82.47	95.50	88.80	87.92	78.67	93.55	**98.52**	96.14	95.18

MRMD	55.10	83.53	91.18	76.67	82.75	65.58	88.71	95.20	92.18	97.59
IB1 Ensembles										
EFSMID+MRMR	**85.09**	**95.15**	**97.10**	94.42	89.72	**95.22**	**96.90**	**96.68**	98.12	**100**
	(0.7)	(0.6)	(0.9)	(0.9)	(0.2)	(0.7)	(0.5)	(0.3)	(1.0)	(0~0.4)
EFSMID+JMI	81.62	93.82	97.08	**95.55**	**91.92**	90.48	95.72	96.62	**100**	**100**
	(0.4)	(0.9)	(0.8)	(0.3)	(0.6)	(0.9)	(0.8)	(0.9)	(1.0)	(0~0.6)
EFSMID+CMIM	78.34	92.60	95.60	95.54	90.27	91.90	95.21	96.62	**100**	**100**
	(0.5)	(0.7)	(0.5)	(0.9)	(0.9)	(0.8)	(0.6)	(0.4)	(1.0)	(0~0.4)
Filters										
MRMR	58.32	80.42	88.92	93.31	84.42	83.62	96.70	91.60	96.64	97.59
JMI	63.36	86.55	96.38	90.13	79.30	80.38	96.62	95.05	95.65	97.53
CMIM	58.35	84.57	96.35	91.10	85.09	86.81	96.61	95.07	95.65	95.18
MRMD	58.33	84.54	91.18	82.22	82.45	85.24	85.48	86.68	92.26	98.80
C4.5 Ensembles										
EFSMID+MRMR	**75.05**	**92.88**	**96.35**	92.23	**89.67**	**84.26**	96.66	**96.65**	**96.15**	**100**
	(0.5)	(0.7)	(0.8)	(0.5)	(0.1)	(0.8)	(0.4)	(0.6)	(1.0)	(0~0.6)
EFSMID+JMI	71.61	90.82	95.78	92.23	87.92	81.90	96.67	95.08	**96.15**	**100**
	(0.8)	(0.7)	(0.9)	(0.2)	(0.3)	(0.9)	(0.7)	(0.9)	(0.8)	(0~0.6)
EFSMID+CMIM	71.63	90.82	95.70	**93.30**	87.98	77.85	**96.90**	95.10	96.08	**100**
	(0.3)	(0.6)	(0.4)	(0.9)	(0.5)	(0.9)	(0.4)	(0.3)	(0.8)	(0~0.8)
Filters										
MRMR	55.09	79.32	94.15	84.47	83.32	75.45	95.14	81.63	83.89	89.16
JMI	56.62	80.42	94.18	78.82	75.27	72.10	95.33	76.62	88.27	92.77
CMIM	58.33	80.43	93.33	80.10	78.10	72.10	95.95	81.12	85.62	93.92
MRMD	53.33	82.47	92.64	80.20	67.24	60.65	88.71	73.33	92.02	83.13

Table 3. Test F1 for microarray datasets. Best F1 for each dataset is highlighted in bold face.

Dataset	1	2	3	4	5	6	7	8	9	10
NBC Ensembles										
EFSMID+MRMR	**85.69**	93.21	91.23	94.12	92.13	94.69	**95.39**	**98.23**	97.56	**100**
EFSMID+JMI	81.54	**93.45**	91.26	**94.36**	91.65	**95.69**	95.21	95.65	**98.32**	**100**
EFSMID+CMIM	83.26	91.56	91.58	93.56	**92.56**	92.21	94.89	97.69	97.32	**100**
Filters										
MRMR	59.63	87.60	91.20	89.52	85.06	77.23	95.21	95.50	98.20	98.82
JMI	69.32	84.50	**91.90**	92.20	83.91	70.60	93.05	98.10	97.70	98.80
CMIM	69.14	87.62	77.90	90.89	86.31	78.52	95.20	98.13	95.60	98.80
MRMD	59.30	84.51	86.60	84.20	82.60	75.70	87.13	95.16	94.23	97.63
SVM Ensembles										
EFSMID+MRMR	**63.21**	**94.58**	96.54	**91.36**	87.65	**81.56**	94.89	95.56	98.25	**100**
EFSMID+JMI	59.65	91.39	95.12	90.56	**90.32**	78.69	**95.89**	94.21	96.85	**100**
EFSMID+CMIM	54.87	92.23	**97.89**	89.36	88.47	78.02	93.65	96.65	**98.32**	**100**
Filters										

MRMR	53.65	82.53	92.45	86.20	86.92	73.57	93.50	95.50	96.30	98.88
JMI	54.56	83.50	90.10	86.90	84.62	73.54	95.12	**98.33**	95.80	97.65
CMIM	54.83	82.50	95.60	86.50	86.68	78.24	93.52	98.20	96.20	95.26
MRMD	51.23	83.51	91.21	72.32	81.42	61.32	88.36	95.20	92.10	96.63
IB1										
Ensembles										
EFSMID+MRMR	**83.25**	**94.78**	97..89	**93.69**	88.69	**95.69**	95.36	**96.32**	98.25	**100**
EFSMID+JMI	80.23	92.68	**97.25**	93.56	**91.20**	90.59	95.21	95.78	**100**	**100**
EFSMID+CMIM	76.59	91.26	95.89	93.60	90.21	91.20	94.78	96.24	**100**	**100**
Filters										
MRMR	57.60	80.20	88.89	92.22	84.10	83.34	**96.88**	91.70	96.25	97.60
JMI	61.23	86.40	95.68	90.90	78.60	80.06	96.16	95.10	95.82	97.62
CMIM	56.20	84.40	96.42	92.10	85.20	87.35	93.50	95.10	95.60	95.20
MRMD	58.20	84.42	91.21	82.73	82.80	85.35	85.72	86.82	92.20	98.82
C4.5										
Ensembles										
EFSMID+MRMR	**73.26**	**91.65**	96.10	91.69	**88.23**	**84.69**	95.36	**96.21**	95.63	**100**
EFSMID+JMI	69.58	90.20	**96.25**	**92.89**	86.96	82.32	**96.80**	94.56	**96.54**	**100**
EFSMID+CMIM	70.26	90.20	96.23	90.58	87.23	78.32	96.30	94.69	95.87	**100**
Filters										
MRMR	54.12	79.40	94.12	83.93	82.65	74.03	95.12	81.32	84.10	88.23
JMI	54.15	80.41	94.12	78.31	74.88	70.60	95.10	77.15	88.30	92.80
CMIM	56.20	80.41	93.41	80.08	77.92	70.60	95.23	81.89	86.13	93.88
MRMD	53.10	82.41	92.60	77.52	67.34	60.25	88.52	73.42	92.10	83.33

Table 4 depicts the average classification accuracy for each method and classifier independent of the datasets. We can see that the best option is to use EFSMID+MRMR combined with NBC classifier. It is also worth noting that even for the rest of the classifiers tested, the ensembles could always achieve the best results, outperforming the results obtained by the filters alone. In general, the proposed framework EFSMID is better than the three feature selection methods.

Table 4. Average of test accuracy for microarray datasets focusing on the classifier.

Classifier	**NBC**	**SVM**	**IB1**	**C4.5**
Ensembles				
EFSMID+MRMR	**95.02**	**90.62**	**94.84**	**91.99**
EFSMID+JMI	94.13	89.49	94.28	90.82
EFSMID+CMIM	94.33	89.49	93.61	90.54
Filters				
MRMR	87.90	86.66	87.15	82.16
JMI	88.09	87.26	88.10	81.04
CMIM	87.92	87.34	88.48	81.90
MRMD	84.89	82.85	84.72	77.37

Table 5 shows the average classification accuracy for each dataset and method independent of the classifier, which could help clarify which method is the best for a given dataset. As we can see, EFSMID+MRMR is significantly better in the maximum number of datasets, followed by EFSMID+CMIM. Adaboost M1 obtains the worst results on the multiple classification datasets, 9 Tumors, BrainTumor, 11 Tumors, NCI and SRBCT, due to the fact that Adaboost M1 is only fit for binary classification dataset. For high-dimensional and small samples microarray data, EFSMID as a different feature spaces ensemble method obtains better classification performance in most cases than the ensemble methods with different training samples, such as Adaboost M1 and Bagging.

Table 5. Average of classification accuracy focusing on the datasets.

Dataset	1	2	3	4	5	6	7	8	9	10
Ensembles										
EFSMID+MRMR	77.82	**94.26**	95.63	93.60	90.21	**88.94**	**96.42**	**96.68**	97.62	**100**
EFSMID+JMI	74.10	92.89	94.97	93.89	**90.74**	86.45	95.72	95.47	97.58	**100**
EFSMID+CMIM	73.34	92.01	95.04	**94.43**	90.13	85.04	95.21	96.66	**98.07**	**100**
Filters										
MRMR	57.59	82.46	91.71	89.17	85.04	77.91	95.11	91.01	93.63	96.08
JMI	62.69	83.76	93.14	87.81	81.06	74.58	95.05	92.08	94.38	96.66
CMIM	60.85	83.79	90.79	87.79	84.35	79.05	95.29	93.22	93.21	95.75
MRMD	56.74	83.77	90.44	80.88	78.80	71.72	87.50	87.70	92.76	94.28
Other ensembles										
Adaboost M1	20.14	75.23	91.90	66.60	28.13	14.73	88.78	90.12	76.11	59.15
Bagging	66.60	93.85	89.72	86.65	85.66	50.82	90.32	83.30	82.09	**100**
Randomforest	**83.30**	92.70	94.86	91.18	88.20	86.52	90.32	93.32	94.13	**100**
RandomSubSpace	70.08	82.47	94.18	82.23	86.70	80.30	87.10	81.65	76.14	96.30
RotationForest	75.12	89.61	**98.50**	87.76	83.66	85.27	91.90	91.69	90.10	97.50

5. Conclusions

In this study, we proposed a novel ensemble feature selection framework EFSMID, which maximizes the relevance between feature subsets and class label as well as the diversity between feature subsets simultaneously. In this framework, features that have more mutual information with class label and more diversity between each other are retained and grouped to different feature subsets by incremental search method. It makes use of the subsets to train base classifiers and then aggregates these classifiers into a consensus outcome by majority voting. To demonstrate the utility of the ensemble method, we compared our method with several state-of-the-art ensemble learning methods and feature selection methods which are also based on information-theoretic criterion through four classifiers on ten high dimensional datasets. The experimental results clearly show that the proposed framework can perform effective and stable ensemble feature selection.

Acknowledgements

The authors would like to acknowledge the assistance provided by National Natural Science Foundation of China (Grant no.61572180 and no.61602283) and Shandong Provincial Natural ScienceFoundation (Grant no. ZR2016FB10).

References

1. T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust Biomarker Identification for Cancer Diagnosis with Ensemble Feature Selection Methods," *Bioinformatics*, vol. 26, no.3, pp. 392-398, 2010.
2. D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based Learning Algorithms," *Machine learning*, vol. 6, no. 1, pp. 37-66, 1991.
3. H. Ahn, H. Moon, M. J. Fazzari, N. Lim, J. J. Chen, and R. L. Kodell, "Classification by Ensembles from Random Partitions of High-Dimensional Data," *Computational Statistics & Data Analysis*, vol. 51, no. 12, pp. 6166-6179, 2007.
4. V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "An Ensemble of Filters and Classifiers for Microarray Data Classification," *Pattern Recognition*, vol. 45, no. 1, pp. 531-539, 2012.
5. V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Data Classification Using an Ensemble of Filters," *Neurocomputing*, vol. 135, no. 135, pp. 13–20, 2014.
6. L. Breiman, "Random Forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.
7. N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods," *Cambridge university press*, 2000.
8. K. W. De Bock, K. Coussement, and D. Van den Poel, "Ensemble Classification Based on Generalized Additive Models," *Computational Statistics & Data Analysis*, vol. 54, no. 6, pp. 1535-1546, 2010.
9. T. G. Dietterich, "Ensemble Methods in Machine Learning," in *International workshop on multiple classifier systems*, pp. 1-15, 2000.
10. F. Fleuret, "Fast Binary Feature Selection with Conditional Mutual Information," *Journal of Machine Learning Research*, vol. 5, pp. 1531-1555, 2004.
11. T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 8, pp. 832-844, 1998.
12. K. B. Irani,"Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning," In *Proceedings of the*

13th International Joint Conference on Artificial Intelligence, pp. 1022-1027, 1993.
13. 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.
14. L. I. Kuncheva and C. J. Whitaker, "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy," *Machine learning*, vol. 51, no. 2, pp. 181-207, 2003.
15. C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe, "A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 9, no. 4, pp. 1106-1119, 2012.
16. H. Liu, L. Liu, and H. Zhang, "Ensemble Gene Selection for Cancer Classification," *Pattern Recognition*, vol. 43, no. 8, pp. 2763-2772, 2010.
17. D. W. Opitz, "Feature Selection for Ensembles," *AAAI/IAAI*, pp. 379-384, 1999.
18. H. Peng, F. Long, and C. Ding, "Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226-1238, 2005.
19. Y. Piao, M. Piao, K. Park, and K. H. Ryu, "An Ensemble Correlation-Based Gene Selection Algorithm for Cancer Classification with Gene Expression Data," *Bioinformatics*, vol. 28, no. 24, pp. 3306-3315, 2012.
20. J. R. Quinlan, "C4. 5: Programs for Machine Learning," *Morgan Kaufmann Publishers Inc*, 1993.
21. M. Reboiro-Jato, F. Díaz, D. Glez-Peña, and F. Fdez-Riverola, "A Novel Ensemble of Classifiers that Use Biological Relevant Gene Sets for Microarray Classification," *Applied Soft Computing*, vol. 17, pp. 117-126, 2014.
22. J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation Forest: A New Classifier Ensemble Method," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 28, no. 10, pp. 1619-1630, 2006.
23. Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust Feature Selection Using Ensemble Feature Selection Techniques," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 313-325, september 2008.
24. N. X. Vinh, J. Epps, and J. Bailey, "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance," *Journal of Machine Learning Research*, vol. 11, pp. 2837-2854, 2010.
25. Witten I H, Frank E, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations," *San Francisco: Morgan Kaufmann Publishers*, 2000.
26. Yang H H and Moody J E, "Data Visualization and Feature Selection: New Algorithms for Nongaussian Data," *in NIPS*, pp. 687-693, 1999.
27. L. Yijing, G. Haixiang, L. Xiao, L. Yanan, and L. Jinling, "Adapted Ensemble Classification Algorithm Based on Multiple Classifier System and Feature Selection for Classifying Multi-Class Imbalanced Data," *Knowledge-Based Systems*, vol. 94, pp. 88-104, 2016.
28. L. Zhang and P. N. Suganthan, "Random Forests with Ensemble of Feature Spaces," *Pattern Recognition*, vol. 47, no. 10, pp. 3429-3437, 2014.
29. Z.H. Zhou, "Ensemble methods: Foundations and Algorithms," *CRC press*, 2012.
30. Q. Zou, J. Zeng, L. Cao, and R. Ji, "A Novel Features Ranking Metric with Application to Scalable Visual and Bioinformatics Data Classification," *Neurocomputing*, vol. 173, pp. 346-354, 2016.

Jie Cai received the Master degree in computer science from Hunan University and is currently a PhD candidate in College of Computer Science and Electronic Engineering, Hunan university. Her research interests include data mining and computational biology.

Jiawei Luo received the PhD degree in computer science from Hunan University in 2008. She is currently a professor in College of Computer Science and Electronic Engineering, Hunan University. She has published about 50 research papers in various international journals and proceedings of conferences. Her research interests include graph theory, data mining, computational biology, and bioinformatics.

Cheng Liang received the PhD degree in computer science from Hunan University in 2015. She was studying at Donnelly Centre, University of Toronto from 2012 to 2014 as a joint PhD student. She is currently an assistant professor in School of Information Science and Engineering, Shandong normal university. Her research interests include data mining and computational biology.

Sheng Yang received the PhD degree in pattern recognition and intelligent system from Shanghai Jiaotong University in 2005. Now, he is working at Hunan University as an associate professor. His current research interests include feature selection, data mining and machine learning.