Hindawi
Complexity
Volume 2021, Article ID 8358921, 12 pages
https://doi.org/10.1155/2021/8358921

Research Article

An Ensemble Classification Method for High-Dimensional Data Using Neighborhood Rough Set

Jing Zhang, Guang Lu, Jiaquan Li, and Chuanwen Li

School of Computer Science and Engineering, Northeastern University, Shenyang 110004, China

Correspondence should be addressed to Chuanwen Li; lichuanwen@mail.neu.edu.cn

Received 5 July 2021; Revised 10 September 2021; Accepted 9 November 2021; Published 24 November 2021

Academic Editor: Hassan Zargarzadeh

Mining useful knowledge from high-dimensional data is a hot research topic. Efficient and effective sample classification and feature selection are challenging tasks due to high dimensionality and small sample size of microarray data. Feature selection is necessary in the process of constructing the model to reduce time and space consumption. Therefore, a feature selection model based on prior knowledge and rough set is proposed. Pathway knowledge is used to select feature subsets, and rough set based on intersection neighborhood is then used to select important feature in each subset, since it can select features without redundancy and deals with numerical features directly. In order to improve the diversity among base classifiers and the efficiency of classification, it is necessary to select part of base classifiers. Classifiers are grouped into several clusters by k-means clustering using the proposed combination distance of Kappa-based diversity and accuracy. The base classifier with the best classification performance in each cluster will be selected to generate the final ensemble model. Experimental results on three *Arabidopsis thaliana* stress response datasets showed that the proposed method achieved better classification performance than existing ensemble models.

1. Introduction

The development of high-throughput sequencing technology has provided researchers with a large amount of microarray data, and extracting valuable information from it has become a hot research topic in bioinformatics [1, 2]. Plants often encounter various stresses at different growth stages throughout their lives, which may lead to inhibition of growth, leaf injury, and plant death. How to predict these stresses will play a very important role in the development of forestry and agriculture. To eliminate negative influences due to these stresses before the appearance of some symptoms, microarray data is used to diagnose and recognize the type of plant stress.

The microarray data has the characteristics of high-dimensionality, small sample, and high redundancy. Traditional classification algorithms for microarray data have problems such as poor classification stability and low accuracy. Therefore, the analysis of such data is required a classification model with strong processing capability. For high-dimensional microarray data, feature selection is a crucial step towards effective and efficient classification [3, 4]. Therefore, high performance methods for feature selection and sample classification have become increasingly important.

Feature selection is an important process in the analysis of high-dimensional data [5, 6]. Rough set theory is a mathematical tool that deals with imprecise, inconsistent, and incomplete problems [7]. The classical rough set theory is employed in attribute reduction problems, and sometimes it requires equivalent relationship among samples. This kind of tolerance relation seems arbitrary and puzzling. In order to deal with this issue, the generalized rough sets theory which substituted binary relation for originally equivalent relationship was proposed [8]. A new binary tolerance relationship intersection neighborhood for processing numerical data was put forward, and it was employed to select features in microarray data [9], which was more flexible for application on data with complex structure. The proposed model uses a rough set model based on the intersection

neighborhood to select features in each dataset. For feature selection of microarray data, these models are usually designed based on a single data source of microarray data. Because of the biological interaction between genes, fusion of existing biological knowledge into the classification model can improve its classification performance.

This paper firstly uses pathway knowledge to make a preliminary selection of features in stress response microarray data and selects important features using intersection neighborhood rough set. Each feature subset contains genes in pathway knowledge unit. Because of the high redundancy of microarray data, only a very small number of genes are related to classification [10], and a gene selection model based on intersection neighborhood rough set (INRS) is used on each feature subset to select important and nonredundant features for subsequent classification work.

The ensemble classification model is constructed by using the information complementarity among base classifiers; thus, it has more stable and accurate classification performance [11, 12]. These methods have become increasingly important because they have better performance than single classifiers in many areas especially for classification problems with complex data structures [9, 13]. Many ensemble learning methods including Boosting [14], Stacking [15], Bagging [16], and Random Subspace Method [17] are proposed. For classification of microarray data, Meng et al. proposed an ensemble classification model using neighborhood system and rough set theory [18]. They aimed at averaging the results of different classifiers [19]. To speed up feature selection process, Meng et al. proposed a parallel feature selection method using MapReduce [20].

In the case of a large number of base classifiers in ensemble model, there will be some redundant classifiers, resulting in poor overall difference. In order to improve the performance of ensemble classification, it is necessary to select base classifiers. Ensemble pruning methods can be roughly divided into four categories: iterative optimization method, ranking method, clustering method, and pattern mining method. In the clustering based technology, Lin et al. proposed a dynamic base classifier selection strategy based on K-means clustering and cyclic sequence [21]. Zhang and Cao proposed an ensemble pruning method based on spectral clustering [22]. Krawczyk used a cluster-based pruning method in the weighted Bagging ensemble classification [23]. However, these methods do not consider the diversity among base classifiers and the classification performance of the classifiers at the same time when calculating the distance between base classifiers.

In this paper, we propose an ensemble classification method for high-dimensional data using neighborhood rough set (ECHDNRS). Since feature distributions of these subsets, which are generated by neighborhood rough set and pathway, are significantly different, diversity among base classifiers trained on these subsets is obvious. Not only is this method a way of generating different training sets for training base classifiers from features, but also it can be regarded as an improvement of the traditional Random Subspace Method. For the purpose of improving the classification performance of the ensemble model and reducing the space and time consumption, k-means clustering is used to select base classifiers. Using k-means instead of passing data to all base classifiers leads to time and space reduction. A new function which combines Kappa-based diversity and accuracy for calculating the distance between two base classifiers is proposed. All base classifiers are grouped into clusters and the base classifier with the best classification accuracy in each cluster is selected to generate the ensemble classification model. Our contributions can be summarized in four aspects:

(1) Use pathway to preselect features.

Pathway is used as a feature preselection tool for ensemble classification. Each pathway contains a group of features that regulate a biological process; thus, the biological significance of features can be assessed. We associate each pathway with its corresponding features included in microarray data to form different feature subsets.

(2) Employ neighborhood rough set to select important features in each pathway unit.

Neighborhood rough set is used to select important features in each pathway unit. In the feature selection, neighborhood rough set makes the classification ability of the pathway unit unchanged. Moreover, it will reduce the training time of the ensemble classification model.

(3) Combine Kappa-based diversity and accuracy to calculate classifier distance.

We consider the diversity among base classifiers and the classification performance of the classifiers at the same time when calculating the distance between base classifiers. A new function which combines Kappa-based diversity and accuracy for calculating the distance between two base classifiers is proposed.

(4) Demonstrate that ECHDNRS can achieve good classification performance.

Extensive experiments demonstrate the good classification performance of ECHDNRS compared with existing schemes.

The rest of the paper is organized as follows: Section 2 describes Framework of ensemble classification. Feature selection based on pathway and INRS is described in Section 3. Our proposed ensemble pruning method based on k-means is described in Section 4. Experiment results and analysis are discussed in Section 5. Finally, conclusion and future work are given in Section 6.

2. Framework of Ensemble Classification

An ensemble classification model generates many classification models for a certain classification problem and results of these models are comprehensively considered [12]. Generally, there are three ways to generate different base classifiers: (1) training base classifiers on training sets with different samples, such as Bagging; (2) training base classifiers on training sets with different features, such as

Random Subspace; (3) using different classification algorithms to train base classifiers based on the same training set [14]. For microarray data with high dimensionality and small sample size, in order to generate base classifiers with significant diversity, it is suitable to employ a model that trains base classifiers on training sets with different features. In this paper, pathway knowledge is used to preselect features; therefore, for the same microarray dataset, different training sets are generated from which we obtain different base classifiers. The procedure of ECHDNRS is shown in Figure 1.

Our ECHDNRS model consists of five steps:

(i) Step 1: Integrate pathway knowledge to generate different feature subsets. Each pathway lists features contained in a specific path. ECHDNRS combines microarray data with the corresponding pathway knowledge to preselect features, to form feature subsets P_i ($i = 1, 2, ..., m$). Since the pathway knowledge is limited, many features included in microarray data have no associated biological knowledge, and they are used to generate unit P_0.

(ii) Step 2: Select important features using intersection neighborhood rough set model (INRS). It is employed to select significant features in each feature unit P_i ($i = 0, 1, 2, ..., m$); then the unit $P_i'(i = 0, 1, 2, ..., m)$ without redundant features is obtained.

(iii) Step 3: All samples are partitioned into training samples, pruning samples, and test samples, as shown in Figure 2. Each training set T_i is the samples set in which all training samples only contain features in P_i'. Then, $\{T_i\}_{i=0}^m$ are used to train base classifiers, and SVM is selected as classification algorithm.

(iv) Step 4: Prune base classifiers using k-means algorithm. Base classifiers are used to classify pruning samples so as to validate the classification performance. Then k-means clustering groups base classifiers into several clusters based on the classification results on pruning samples. The base classifier with the best classification performance in each cluster will be selected to generate the final ensemble model.

(v) Step 5: Integrate classification results of these selected base classifiers. Each test sample is classified by all selected base classifiers; then, the model integrates the results of different classifiers by majority vote method.

These steps of the ECHDNRS are described in detail in the following subsections.

3. Feature Selection

The performance of classifier depends on the interrelationships among the number of samples, dimension of features, and complexity of classifier [13]. If the number of samples in training set is much smaller than the number of features, it will lead to poor classification performance due to the overfitting of classifier on training set [24]. This behavior is referred to as the peaking phenomenon [25, 26].

In practice, sample number is very small relative to the dimension of feature, usually tens of thousands in microarray data. For the objective of improving the classification performance, feature selection is essential. Therefore, pathway knowledge is used in preselection which is used to generate a number of feature subsets. Moreover, INRS is employed to select important features in each unit.

3.1. Combining High-Dimensional Data with Prior Knowledge. We use high-dimensional microarray data combined with pathway knowledge to generate different subsets. The pathway biological knowledge is downloaded from https://www.arabidopsis.org/biocyc, which was derived from KEGG by Kanehisa [27]. The KEGG pathway database integrates current knowledge on molecular interaction networks, including graphical cellular biochemical processes such as metabolism, cell cycle, signal transduction, membrane transport, and conservative subchannel information. It is a collection of hand-painted metabolic pathways, containing the following aspects of intermolecular interactions and response network: (1) metabolism; (2) genetic information processing; (3) environmental information processing; (4) cellular processes; (5) biological system; (6) human diseases; (7) drug development.

For classification of microarray data, traditional classification models are usually designed based on a single data source of microarray data. Because of the biological interaction between genes, fusion of existing biological knowledge into the classification model can improve its classification performance. Gene Ontology (GO) knowledge was first applied to cancer prediction. Related experiments show that combining biological knowledge can improve the accuracy of prediction results and enhance its biological interpretability and credibility [28]. After that, the prediction model combined with pathway knowledge was also applied to the prediction of cancer [29]. In recent years, a classification model at the pathway level combined with the superbox principle was applied to disease classification [30].

The proposed ECHDNRS model eliminates the randomness of the traditional Random Subspace, without using prior knowledge; it randomly extracts features to form feature subsets. Feature selection integrated with biological knowledge for plant stress response improves the biological interpretation of the results [18]. Three examples of pathway are shown in Figure 3, where p_{ij} represents the jth feature contained in pathway i. For pathway units, the number of features ranges from 1 to more than 200.

There exist features that are contained in microarray data but without corresponding pathway. Wilcoxon rank sum is employed in preselection for features that are not associated with any pathway annotation, and then 200 top-ranked features are used to generate unit P_0. Wilcoxon rank sum test is suitable for ranking of samples that do not meet specific probability distribution such as Gaussian distribution and it is suitable for binary classification samples.

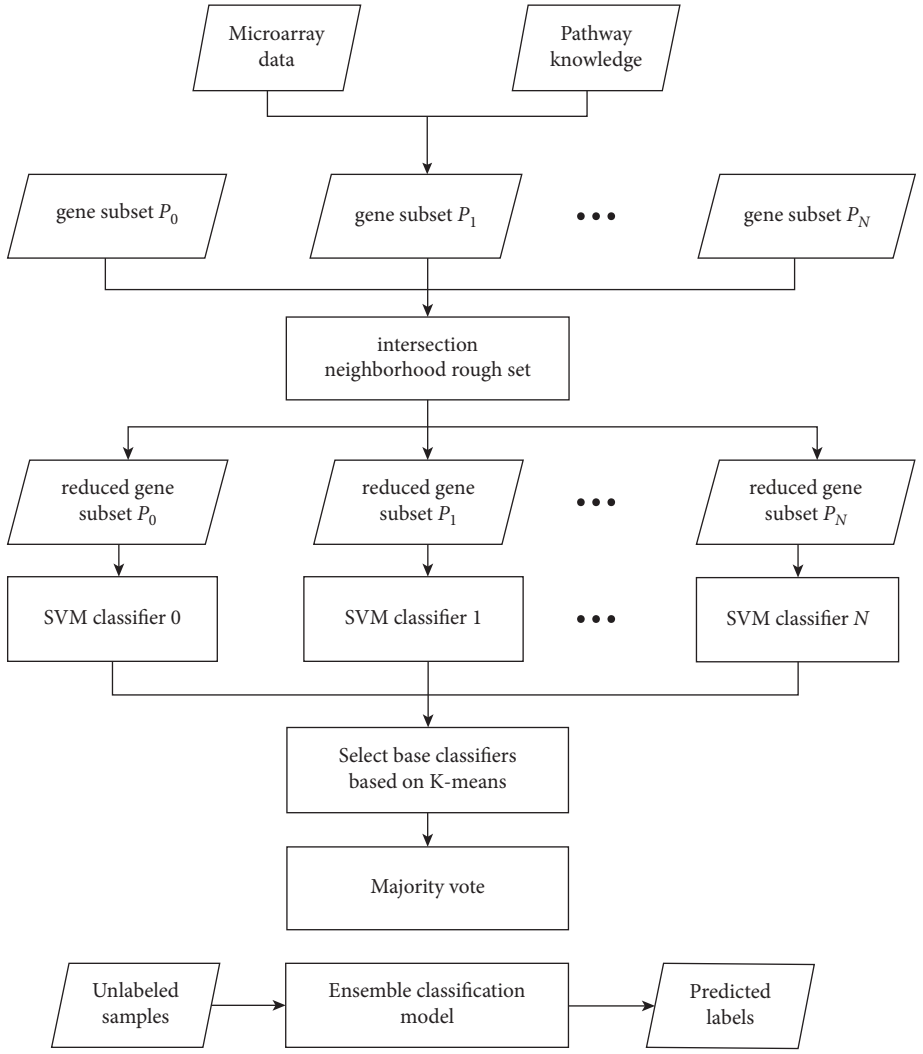

Figure 1: Flowchart of the proposed model ECHDNRS.

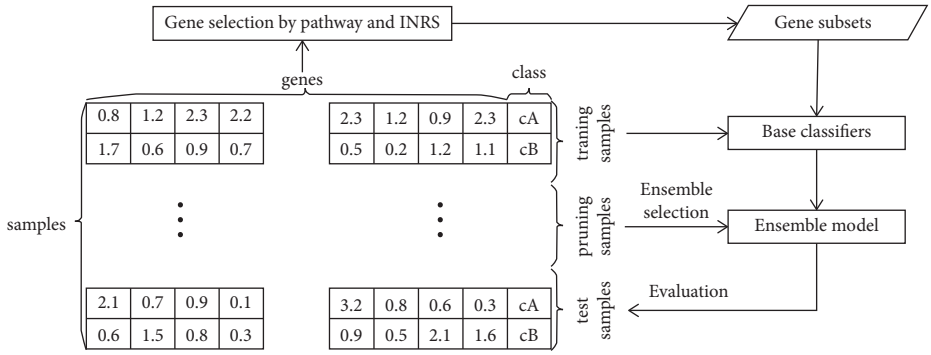

Figure 2: An overview of classifying microarray data.

For each feature, the expression quantity of every sample in microarray data is viewed as observations in statistical test. Thus, each feature has two groups of observations depending on the class labels of the sample, respectively denoted by $X = \{x_i \mid i = 1, 2, \ldots, n_1\}$ which represent expression quantity corresponding to samples belonging to class 1 and $Y = \{y_j \mid j = 1, 2, \ldots, n_2\}$ which represent expression quantity correspond to samples belong to class 2, where n_1 and n_2 are the number of samples belonging to class 1 and class 2, respectively. All of the $(n_1 + n_2)$ samples are ranked based on the expression quantity in ascending order. Since it is possible that many samples have the same expression quantity, they need to be adjusted, so as to obtain the same rank which is the average of all the ranks. For two

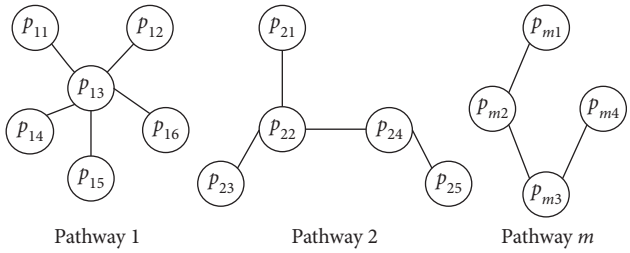

Figure 3: Schematic diagram of pathway.

groups of observations $X = \{x_i \mid i = 1, 2, \ldots, n_1\}$ and $Y = \{y_j \mid j = 1, 2, \ldots, n_2\}$, when $n_1, n_2 > 10$, the test statistic of Wilcoxon rank sum test for each feature is defined as follows:

$$Z = \frac{U - n_1 n_2/2}{\sqrt{[n_1 n_2 (n_1 + n_2 + 1)]/12}} \sim N(0, 1). \tag{1}$$

U is the smaller one of $U_1 = n_1 n_2 + [n_1(n_1 + 1)] - T_1$ and $U_2 = n_1 n_2 + [n_2(n_2 + 1)] - T_2$. T_1 and T_2 are the sum of rank of class 1 and class 2, respectively. The test statistic Z obeys standard normal Gaussian distribution with mean 0 and variance 1. The P-value for feature g is as follows:

$$p = P\left(|Z| > \left|Z_g\right|\right), \tag{2}$$

whereby Z_g is the test statistic Z based on Wilcoxon rank, the sum test for feature g, and $P(|Z| > |Z_g|)$ represents the probability of $|Z| > |Z_g|$. The smaller p is, the bigger the diversity of feature g in two classes is. Finally, all features are ranked based on p in descending order, and then top-N features are selected as the output of feature preselection.

By combining microarray data with pathway knowledge, information table is obtained and shown in Table 1.In the table, P_0 contains features which are not associated with any pathway annotations and are preselected by Wilcoxon rank sum test. P_i ($i = 1, 2, ..., m$) is a biological knowledge unit including features of a pathway corresponding to microarray data, and m is the number of pathways selected as prior knowledge. For a feature $p_{ij} \in P_i$ ($j = 1, 2, ..., |P_i|$) and a sample s_k ($k = 1, 2, ..., n$), the value v_{ijk} is the expression quantity of feature p_{ij} in sample s_k. There exist some redundant features in some pathway for classification, so further feature selection is needed. After the process of forming these units, feature selection model based on INRS is employed to delete redundant features in each unit.

3.2. Feature Selection Based on Intersection Neighborhood Rough Set Model. For $GEDT = \{S, P_i \cup D, V, f\}$, $S(s_1, s_2, \ldots, s_n)$ is the microarray sample set, P_i represents the unit of pathway knowledge which contains related features, and $D = \{d\}$ is the class label. For each feature $p_{ij} \in P_i$ and sample $s_k \in S$, v_{ijk} is the expression value of feature p_{ij} for sample s_k, and d_k is the class label of s_k.

R is binary relation on data space U for objects x and y. If y is included in the neighborhood of x, the neighborhood of x is defined as follows [31]:

$$N_R(x) = \{y \mid xRy, y \in U\}. \tag{3}$$

Since all the gene expression microarray data is numerical, we focus on the tolerance relation defined for numerical features. We use the intersection neighborhood as the binary relation. For object $x \in U$, based on P_i the relation is defined in [7, 18]

$$N_{P_i}(x) = \left\{y \mid \forall p_{ij} \in P_i, \left|f_{p_{ij}}(x) - f_{p_{ij}}(y)\right| \le \delta_{p_{ij}}, y \in U\right\}. \tag{4}$$

In order to simplify the process of feature selection, we assume that each feature has the same threshold δ. For objects x and y, when the distance between each feature $p_{ij} \in P_i$ is less than or equal to δ, then y is in the neighborhood based on P_i of x.

Based on the above binary relation, for a subset of objects $X \subseteq U$, the definitions of lower and upper approximations in the extended rough set theory are defined in (3) and (4), respectively [31]:

$$\begin{aligned} \underline{apr}_{P_i}(X) &= \left\{x | N_{P_i}(x) \subseteq X, x \in U\right\}, \\ \overline{apr}_{P_i}(X) &= \left\{x | N_{P_i}(x) \cap X \neq \Phi, x \in U\right\}. \end{aligned} \tag{5}$$

The definitions of positive, negative, and boundary regions based on intersection neighborhood relation are as follows [31]:

$$\begin{aligned} POS_{P_i}(X) &= \underline{apr}_{P_i}, \\ NEG_{P_i}(X) &= U - \overline{apr}_{P_i}, \\ BN_{P_i}(X) &= \overline{apr}_{P_i} - \underline{apr}_{P_i}. \end{aligned} \tag{6}$$

The principle of feature selection model based on the rough set is to keep classification ability unchanged. The classification ability in rough set is defined as the number of training samples included in the positive region based on decision feature; for each unit P_i, it is represented by $POS_{RE\,D}(D) = POS_{P_i}(D)$. It is a kind of NP-hard problem to find an optimal subset of features. Features in each unit P_i are ranked using Wilcoxon rank sum test. Therefore, the proposed feature selection model takes backward strategy in which the rank of features is employed as heuristic information. Every feature ranked from top to bottom in the unit is evaluated such that when deleting a feature from the feature set, if it meets the condition of unchanging the classification ability, it is removed from the unit; otherwise, it cannot be removed. This method can maintain features which have better classification ability, and it can also include fewer features in the selected feature subsets. The feature selection algorithm based on the INRS is described in Algorithm 1.

The computational complexity of the intersection rough set-based algorithm is $O(|Pi|^2|S|^2)$, whereby $|P_i|$ is the number of features contained in P_i and $|S|$ is the number of samples. This feature selection method takes full advantage of global information of each feature subset P_i.

Table 1: Information table formed by combing microarray data with pathway knowledge.

Sample	P_0			P_1			P_2			...	P_m		
	p_{01}	p_{02}	...	p_{11}	p_{12}	...	p_{21}	p_{22}	...		p_{m1}	p_{m2}	...
s_1	v_{011}	v_{021}	...	v_{111}	v_{121}	...	v_{211}	v_{221}	...		v_{m11}	v_{m21}	...
s_2	v_{012}	v_{022}	...	v_{112}	v_{122}	...	v_{212}	v_{222}	...		v_{m12}	v_{m22}	...
...													
s_n	v_{01n}	v_{02n}	...	v_{11n}	v_{12n}	...	v_{21n}	v_{22n}	...		v_{m1n}	v_{m2n}	...

Input: $GEDT = \{S, P_i \cup D, V, f\}$
δ //the list of intersection neighborhood threshold
Output: RED // a set of features which is a reduction of unit P_i
Step 1: For each sample $s_k \in S$, calculate the intersection neighborhood $N_{P_i}(s_k)$ based on different threshold δ
Step 2: Divide the sample set S based on the class label $D = \{d\}$ to obtain the equivalence classes which are represented as $S/IND(D)$ (where samples with the same class label contained in one equivalence class).
Step 3: Calculate the positive region defined on intersection neighborhood $POS_{P_i}(D) = \cup_{X \in S/IN\ D(D)} POS_{P_i}(X)$ based on all of the features in P_i.
Step 4: Start with $RED = P_i$.
Step 5: As Step 1 for sample $s_k \in S$, calculate the intersection neighborhood $N_{RE\ D-\{p_{ij}\}}(s_k)$. Then as Step 3 calculate the positive region based on RED-$\{p_{ij}\}$, $POS_{RE\ D-\{p_{ij}\}}(D) = \cup_{X \in S/IN\ D(D)} POS_{RE\ D-\{p_{ij}\}}(X)$;
If $POS_{RE\ D-\{p_{ij}\}}(D) = POS_{P_i}(D)$, then make $RED = RED - \{p_{ij}\}$.
Step 6: Repeat Step 5 until all the features p_{ij} in the subset P_i are validated, then use final RED as a reduction of P_i, marked as $P_i' = RED$.

Algorithm 1: Feature selection.

3.3. Ensemble Pruning Based on Clustering. For ensemble classification, many base classifiers are generated for the same problem, a large amount of memory and considerable computational cost are needed [32]. Therefore, classifier pruning is essential to ensemble model. Additionally, Zhou revealed that the ensemble of a proper subset of base classifiers sometimes outperforms the original ensemble [33, 34].

3.3.1. Classifier Distance Based on Diversity and Accuracy. In order to improve the performance of the ensemble classifier, classifiers with significant diversity are selected. Diversity can be viewed as a measure of dependence, complement, or orthogonality among classifiers [35]. Diverse classifier ensembles are preferred. There exist many approaches to measure diversity among binary classifier output including the Q statistic, the correlation, the disagreement, the double fault, the entropy of the votes, the difficulty index, the Kohavi-Wolpert variance, the interrater agreement, and the generalized diversity [36].

Cohen proposed Kappa statistic as the index for consistency judgment. In practice, it can measure the consistency of diagnosis well; therefore, it has been widely used in clinical trials. Kappa is also used to evaluate the classification performance of base classifier. This index compensates for classifications that may be due to randomness. It is considered as a standard statistically robust metric for measuring the accuracy in multiclass problems [37].

In our method, diversity is measured based on Kappa coefficient. The evaluation method implemented in the proposed method does not compute the Kappa index as a global performance measure for each candidate classifier. We calculate the specific Kappa value for the similarity between two candidate classifiers (kappa by similarity). Kappa coefficient of the output of two base classifiers is calculated as follows [38]:

$$Kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

$$\Pr(a) = \frac{C^{tt}}{N} + \frac{C^{ff}}{N},$$

$$\Pr(e) = \frac{C^{tt} + C^{tf}}{N} \times \frac{C^{tt} + C^{ft}}{N} + \frac{C^{ft} + C^{ff}}{N} \times \frac{C^{tf} + C^{ff}}{N}. \tag{7}$$

There are two base classifiers h_i and h_j which are the i-th and the j-th classifiers, respectively. N is the total number of samples. C^{tt} represents the number of samples which were correctly classified by h_i and h_j; on the contrary, C^{ff} stands for the number of samples that were incorrectly classified by h_i and h_j; C^{tf} is the number of samples that were correctly classified by h_i but were incorrectly classified by h_j, and C^{ft} is the number of samples that were incorrectly classified by h_i but were correctly classified by h_j.

The value of Kappa ranges from −1 to 1, and *kappa* stands for the value of Kappa coefficient. When $kappa < 0$, it means that the consistency is worse due to randomness; when $kappa > 0$, it means that the bigger the value, the better the consistency. We define the diversity distance between two base classifiers as follows: when $kappa \leq 0$, the distance is 1; when $kappa > 0$, the distance is $1 - kappa$. The diversity

distance is symmetric for each pair of classifiers, since $D(d)_{ij} = D(d)_{ji}$.

$$D(d)_{ij} = \begin{cases} 1, & kappa \le 0, \\ 1 - kappa, & kappa > 0. \end{cases} \tag{8}$$

Giacinto and Roli stated that ensemble classifiers should be accurate and diverse [39]. Thus, when considering the diversity between two base classifiers, it is also essential to consider their classification accuracy. Suppose that there are m samples in pruning set; c_{ik} denotes the actual output of the i-th classifier on the k-th sample; when $c_{ik} = 0$, the k-th sample is correctly classified by the i-th classifier; otherwise, it represents the sample that is incorrectly classified. If the actual output c_{ik} is 0 as well as $c_{jk} = 0$, then $c_{ik}c_{jk} = 1$; otherwise, $c_{ik}c_{jk} = 0$. Accuracy distance between them is defined as follows:

$$D(a)_{ij} = \begin{cases} 1 - \dfrac{1}{m}\displaystyle\sum_{k=1}^{m} c_{ik}c_{jk}, & i \ne j, \\ 0, & i = j. \end{cases} \tag{9}$$

We consider both of the diversity distance and accuracy distance in one distance function and define the distance between base classifiers h_i and h_j as follows:

$$D_{ij} = \alpha D(d)_{ij} + (1 - \alpha)D(a)_{ij}, \tag{10}$$

where α[0, 1] is the weight for diversity distance, D_{ij}[0, 1] is also symmetric and nonnegative, and D_{ii} equals 0 for each I; therefore, it meets the requirements of the distance definition.

3.4. Pruning Base Classifiers Using K-Means. An ensemble pruning method based on k-means clustering is proposed which considers both Kappa-based diversity and accuracy of classifiers. The objective of clustering is to partition base classifiers into many homogenous clusters in which classifiers within a cluster are more similar to each other than those belonging to different clusters. This means that classifiers belonging to different clusters are more diverse. Then an exemplar is selected from each cluster to participate in the ensemble model.

An improved k-means clustering is used to partition the set of base classifiers $H = \{h_1, h_2, ..., h_N\}$ into k clusters based on the distance we defined. Firstly, k cluster centroids labeled as $\{C_j\}_{j=1}^{k}$ are randomly selected from all the base classifiers. Secondly, calculate the distance of each classifier between it and every centroid; then it belongs to the cluster of the nearest centroid. Thirdly, in normal k-means clustering, in order to adjust the centroids, the average of all members in one cluster is viewed as the calculated centroid. However, in this situation, these calculated centroids may not stand for the real base classifiers. For each calculated centroid, the original method is improved by selecting one classifier with the smallest distance between them as the new centroid. Finally, repeat the above operations until reaching the iteration that gives the best output or until centroids of all base classifiers are unchanged. To obtain the optimal number of clusters, the number of clusters k is increased gradually until $\min\sum_i^N D(h_i, C_{h_i})$ starts to deteriorate. When the optimal number k of the clusters is obtained, according to the assumption in [40], the agreement among the classifiers from the same cluster is large, so the majority of the classifiers can be removed. Then the classifier with the best classification performance in each cluster is selected as the exemplar to participate in the ensemble model.

We use a group of selected classifiers to classify test samples and integrate the results by majority vote method. Since among the selected base classifiers there exists significant diversity, if most of the base classifiers are consistent, then the result will have a higher credibility. If the differences among the base classifiers are not obvious, in some cases it is possible that most of the classifiers misclassify the samples, and then the ensemble classifier will also incorrectly classify the samples.

Training time includes feature selection time, base classifiers generation time, and classifier pruning time. The feature selection based intersection neighborhood rough set is very time consuming, since it needs to calculate positive region for each feature unit. Base classifiers generation time is related to classification algorithm. Classifier pruning is to reduce the classification time; thus, it is necessary for ensemble classification. After the training process, inference time associates with the base classifier number selected by the ensemble in our proposed model. Therefore, ensemble pruning based on k-means can reduce the inference time.

4. Experimental Results and Discussion

4.1. Dataset and Experiment Settings. *Arabidopsis thaliana* is usually used to study the responses of plants to different types of stress [41], because of its abundant biological experiment data and information encoded in gene annotations. In this paper, three plant stress response datasets about *Arabidopsis thaliana* and corresponding pathway knowledge are applied in the experiment to test the performance of the proposed ECHDNRS model. The datasets are Arabidopsis-Drought, *Arabidopsis*-Oxygen, and *Arabidopsis*-TEV, which are responses to drought, oxygen, and *Potyvirus* (TEV) stress, respectively. All the three datasets can be downloaded from GEO (Gene Expression Omnibus) website (http://www.ncbi.nlm.nih.gov/geo/). Each dataset has two classes. Detailed information about the three *Arabidopsis thaliana* datasets is shown in Table 2, and the experimental group and the control group are denoted by Class A and Class B, respectively.

The data values are normalized into the range of [−1, 1] before intersection neighborhoods of objects are constructed, to eliminate the deviation of the classification result caused by different property ranges. In order to observe the impact of the change in δ threshold of intersection neighborhood on the classification performance, δ is set ranging from 0.05 to 0.95 with a difference of 0.1. The weight of Kappa-based diversity distance α in the process of ensemble pruning ranges from 0.1 to 0.9 with a step of 0.1. For all samples, 60% of them are used as training samples, 20% as pruning samples, and the rest as testing samples. The number of samples is limited in microarray data, and in

Table 2: Detailed information about the three *Arabidopsis thaliana* datasets.

Dataset	Feature	Sample	Class *A*	Class *B*
Arabidopsis-Drought	20728	59	29	30
Arabidopsis-Oxygen	22810	79	43	36
Arabidopsis-TEV	32916	53	28	25

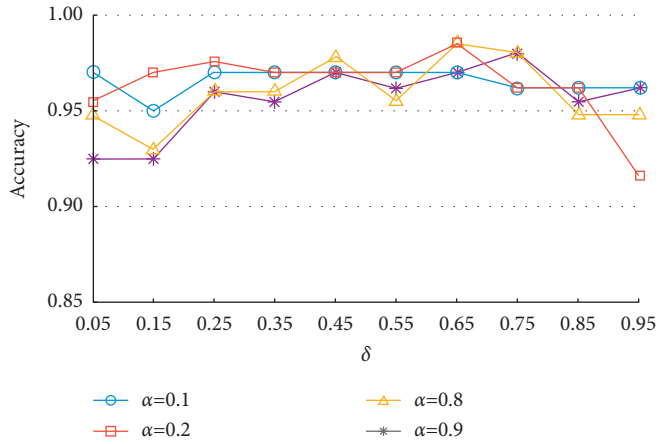

Figure 4: Average accuracies of ECHDNRS on *Arabidopsis*-Drought dataset.

order to eliminate the coincidence caused by random sampling, the average classification performance of 10 times sampling is used as the final result. In this paper, SVM is used to classify the microarray data. We set the kernel function of SVM classifier as RBF function ($K(x, y) = exp(-\gamma||x-y||^2)$) which has a strong ability to adapt to different datasets and use libSVM to implement SVM.

4.2. Experiment Result Analysis. The classification accuracies of method ECHDNRS are shown in Figures 4–6. For three datasets, when α is set to 0.1, 0.2, 0.8, and 0.9, the ensemble model achieves better performance. For *Arabidopsis*-Drought, *Arabidopsis*-Oxygen, and *Arabidopsis*-TEV dataset, the ensemble model achieves the best performance when δ is set to 0.65, 0.55, and 0.55, respectively. When δ is smaller than peak-value, the performance is more stable and better than that when δ is bigger than the peak-value.

Feature selection, training, classifier pruning, ensemble classification without pruning, and pruning ensemble classification time are shown in Table 3. The feature selection process is very time consuming; it takes at least 1820s. On these datasets, classifier pruning reduces the classification time and improves classification performance; thus, it is necessary for ensemble classification.

Hence, ECHDNRS selects many feature units. The average p-value of these selected features on three datasets are shown in Table 4. ECHDNRS makes the ensemble model obtain a good classification performance, and some of the selected features are not performed very well on p-value.

The sum of distance between all base classifiers and its centroid is shown in Figure 7. When k is bigger than 7, it begins to deteriorate. Therefore, we set k to 7.

The evaluation criterion of binary classification is based on four simple criteria: True Positives (*TP*), False Positives (*FP*), True Negatives (*TN*), and False Negatives (*FN*). In this paper, four evaluation criteria are used to evaluate the comparison results; they include accuracy (*ACC*), sensitivity (*SN*), specificity (*SP*), and geometric mean (*G-mean*). They are defined as follows:

$$\begin{aligned} ACC &= \frac{TP + TN}{(TP + TN + FP + FN)}, \\ SN &= \frac{TP}{(TP + FN)}, \\ SP &= \frac{TN}{(TN + FP)}, \\ G - mean &= \sqrt{SN * SP}. \end{aligned} \tag{11}$$

ACC evaluates the classification accuracy of all the samples. *SN* and *SP* measure the classification accuracies of the samples belonging to positive and negative classes, respectively. *G-mean* comprehensively evaluates the classification ability of positive and negative classes.

The comparison methods include four classical ensemble models and one single model; they are Random Subspace, Bagging, AdaboostM1, Stacking, and SVM; they are implemented in Weka [42]. Two single distances, diversity distance based on Kappa and accuracy distance for clustering base classifiers, are compared with ECHDNRS; they are named as DECHDNRS and AECHDNRS, respectively. The accuracy distance is a general calculation formula of the distance between the base classifier. Kappa coefficient is a common evaluation index for the performance of classifier. Therefore, ensemble pruning methods, which are based on

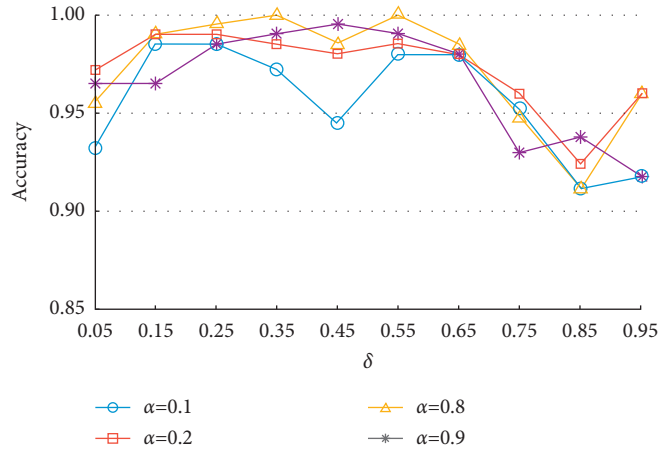

Figure 5: Average accuracies of ECHDNRS on *Arabidopsis*-Oxygen dataset.

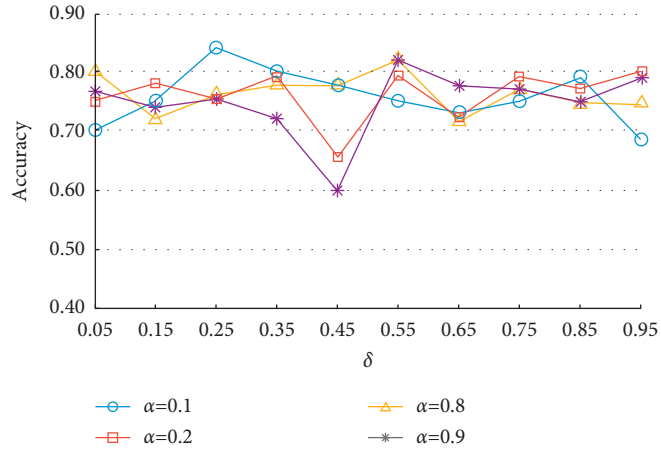

Figure 6: Average accuracies of ECHDNRS on *Arabidopsis*-TEV dataset.

Table 3: Runtime on three datasets.

Datasets	Feature selection	Training	Pruning	Ensemble	Pruning ensemble
Arabidopsis-drought	1820.4645	0.1344	0.0184	0.0099	0.0006
Arabidopsis-oxygen	4450.6792	0.1343	0.0188	0.0284	0.0016
Arabidopsis-TEV	1824.7600	0.0945	0.0618	0.0125	0.0007

Table 4: Average p-value of selected features.

Datasets	p-value
Arabidopsis-drought	0.5164
Arabidopsis-oxygen	0.3765
Arabidopsis-TEV	0.2043

them, are also compared with ECHDNRS. SVM is used as the base classifiers for all ensemble models. All of the methods based on ensemble pruning employ the same method as ECHDNRS to generate base classifiers. The intersection neighborhood threshold δ has different effect on the performance of different ensemble models; therefore, the average classification performance of different δ is used to compare with other methods. The comparison results of ECHDNRS with other ensemble methods are shown in Tables 5–7.

For each dataset, for example, we rank these nine classification methods according to the classification accuracies: the classification method with the best accuracy is ranked the 1st, and the method with the worst accuracy is ranked 9th. Then, for each classification method, calculate the average ranking over the three datasets. The average rankings of three datasets are listed in Table 8.

From the performance of all methods on three datasets and the rankings of them, the ECHDNRS obtains the best performance on *SP* and *G-mean*, and it is similar to DECHDNRS on accuracy. On *SN*, ECHDNRS is worse than other methods, but it balances the classification ability of positive and negative classes in a better way. Since it can

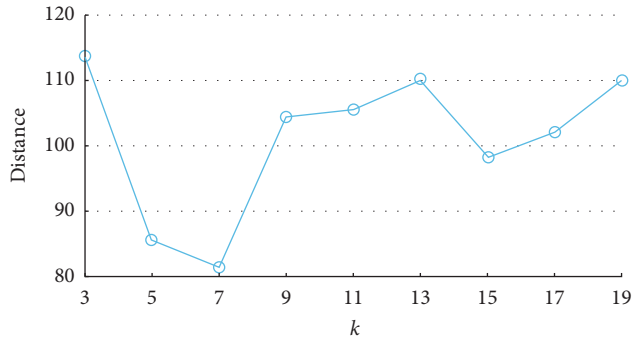

Figure 7: Distance between classifiers and centroids.

Table 5: Classification results on *Arabidopsis*-Drought dataset.

Ensemble	*ACC* (%)	*SN* (%)	*SP* (%)	*G-mean*
SVM	91.67	100.00	83.33	0.9129
Random	91.67	100.00	83.33	0.9129
Bagging	87.50	100.00	75.00	0.8660
AdaboostM1	95.83	100.00	91.67	0.9574
Stacking	91.67	100.00	83.33	0.9129
Kappa	93.54	94.67	92.57	0.9360
DECHDNRS	96.15	97.33	95.14	0.9623
AECHDNRS	95.77	97.67	94.14	0.9587
ECHDNRS	96.08	96.71	95.54	0.9611

Table 6: Classification results on *Arabidopsis*-Oxygen dataset.

Ensemble	*ACC* (%)	*SN* (%)	*SP* (%)	*G-mean*
SVM	87.50	100.00	66.67	0.8165
Random	87.50	100.00	71.44	0.8452
Bagging	96.88	100.00	91.67	0.9574
AdaboostM1	96.88	100.00	91.67	0.9574
Stacking	87.50	100.00	66.67	0.8165
Kappa	96.18	98.80	92.87	0.9570
DECHDNRS	96.67	98.60	91.75	0.9505
AECHDNRS	96.28	98.30	93.13	0.9565
ECHDNRS	96.61	99.13	93.44	0.9619

Table 7: Classification results on *Arabidopsis*-TEV dataset.

Ensemble	*ACC* (%)	*SN* (%)	*SP* (%)	*G-mean*
SVM	33.33	100.00	0.00	0.0000
Random	33.33	100.00	0.00	0.0000
Bagging	33.33	100.00	0.00	0.0000
AdaboostM1	52.38	85.71	35.71	0.5533
Stacking	33.33	100.00	0.00	0.0000
Kappa	66.83	75.00	55.40	0.6422
DECHDNRS	75.58	80.86	68.20	0.7415
AECHDNRS	74.42	78.43	68.80	0.7331
ECHDNRS	75.65	82.21	66.45	0.7378

Table 8: Average ranking on three datasets.

Ensemble	*ACC*	*SN*	SP	*G-mean*
SVM	6.33	1.00	6.67	6.67
Random	7.00	1.00	6.67	6.67
Bagging	5.33	1.00	6.67	5.67
AdaboostM1	3.00	2.33	5.00	3.67
Stacking	7.67	1.00	7.67	7.67
Kappa	5.00	8.33	3.67	4.33
DECHDNRS	2.00	7.33	2.67	2.67
AECHDNRS	4.00	7.67	2.00	3.67
ECHDNRS	2.33	6.67	1.67	1.67

classify the samples belonging to negative class well, other methods perform much worse on them. Therefore, it obtains better performance on *G-mean*, which comprehensively evaluates the classification ability of positive and negative classes. Generally, these methods based on ensemble pruning perform better than classical ensemble models.

5. Conclusions

An ensemble classification method ECHDNRS for plant stress response is proposed in this paper. Combining microarray data with pathway knowledge to eliminate randomness of the traditional Random Subspace, then the feature selection model based on intersection neighborhood rough set reduces redundant features in each feature unit. Furthermore, in order to improve the classification performance of the ensemble model, the hybrid approach of classification and clustering is employed to select base classifiers. The *k*-means clustering algorithm that employs the proposed distance function which is the combination of Kappa-based diversity and accuracy groups all base classifies into several clusters, and the base classifier with the best classification accuracy in each cluster is selected. Experimental results on three *Arabidopsis thaliana* stress-related datasets show that the proposed method obtains better results than classical ensemble methods including Random Subspace, Bagging, AdaboostM1, and Stacking, and it also performs better than traditional Kappa pruning and clustering pruning methods based on single distance. How to reduce the time consumption when performing feature selection based on intersection neighborhood rough set is the topic of our future work.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61872071) and the Fundamental Research Funds for the Central Universities (N2116010).

References

[1] P. Maji and P. Garai, "On fuzzy-rough attribute selection: criteria of max-dependency, max-relevance, min-redundancy, and max-significance," *Applied Soft Computing*, vol. 13, no. 9, pp. 3968–3980, 2013.

[2] P. Maji and S. Paul, "Rough-fuzzy clustering for grouping functionally similar genes from microarray data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 2, pp. 286–299, 2012.

[3] Z. Y. Algamal and M. H. Lee, "Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification," *Computers in Biology and Medicine*, vol. 67, pp. 136–145, 2015.

[4] D. Liu, T. Li, and D. Liang, "Incorporating logistic regression to decision-theoretic rough sets for classifications," *International Journal of Approximate Reasoning*, vol. 55, no. 1, pp. 197–210, 2014.

[5] W. Ma, X. Zhou, H. Zhu, L. Li, and L. Jiao, "A two-stage hybrid ant colony optimization for high-dimensional feature selection," *Pattern Recognition*, vol. 116, Article ID 107933, 2021.

[6] O. A. Alomari, S. N. Makhadmeh, M. A. Al-Betar et al., "Gene selection for microarray data classification based on Gray Wolf Optimizer enhanced with TRIZ-inspired operators," *Knowledge-Based Systems*, vol. 223, Article ID 107034, 2021.

[7] Z. a. Pawlak, "Rough sets," *International Journal of Computer & Information Sciences*, vol. 11, no. 5, pp. 341–356, 1982.

[8] Y. Y. Yao and T. Y. Lin, "Generalization of rough sets using modal logics," *Intelligent Automation & Soft Computing*, vol. 2, no. 2, pp. 103–119, 1996.

[9] J. Meng, J. Zhang, R. Li, and Y. Luan, "Gene selection using rough set based on neighborhood for the analysis of plant stress response," *Applied Soft Computing*, vol. 25, pp. 51–63, 2014.

[10] L. J. Zhang and Z. J. Li, "Gene selection for cancer classification in microarray data," *Journal of Computer Research and Development*, vol. 46, no. 5, pp. 794–802, 2009.

[11] M. Landesfeind, A. Kaever, K. Feussner et al., "Integrative study ofArabidopsis thalianametabolomic and transcriptomic data with the interactive MarVis-Graph software," *PeerJ*, vol. 2, p. e239, 2014.

[12] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1/2, pp. 1–39, 2010.

[13] A. K. Jain, P. W. Duin, and J. Jianchang Mao, "Statistical pattern recognition: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[14] R. E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, no. 2, pp. 197–227, 1990.

[15] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.
[16] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
[17] T. K. Tin Kam 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.
[18] J. Meng, J. Zhang, and Y. Luan, "Gene selection integrated with biological knowledge for plant stress response using neighborhood system and rough set theory," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 2, pp. 433–444, 2015.
[19] L. Nanni and A. Lumini, "Using ensemble of classifiers in bioinformatics," *Machine learning research progress*, vol. 36, pp. 3896–3900, 2008.
[20] J. Meng, J. Zhang, Y.-S. Luan, X.-Y. He, L.-S. Li, and Y.-F. Zhu, "Parallel gene selection and dynamic ensemble pruning based on Affinity Propagation," *Computers in Biology and Medicine*, vol. 87, pp. 8–21, 2017.
[21] C. Lin, W. Chen, C. Qiu, Y. Wu, S. Krishnan, and Q. Zou, "LibD3C: ensemble classifiers with a clustering and dynamic selection strategy," *Neurocomputing*, vol. 123, pp. 424–435, 2014.
[22] H. Zhang and L. Cao, "A spectral clustering based ensemble pruning approach," *Neurocomputing*, vol. 139, pp. 289–297, 2014.
[23] B. Krawczyk, "Forming ensembles of soft one-class classifiers with weighted bagging," *New Generation Computing*, vol. 33, no. 4, pp. 449–466, 2015.
[24] A. K. Jain and B. Chandrasekaran, "39 Dimensionality and sample size considerations in pattern recognition practice," *Handbook of Statistics*, vol. 2, pp. 835–855, 1982.
[25] S. Raudys and V. Pikelis, "On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-2, no. 3, pp. 242–252, 1980.
[26] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 3, pp. 252–264, 1991.
[27] M. Kanehisa, S. Goto, M. Hattori et al., "From genomics to chemical genomics: new developments in KEGG," *Nucleic Acids Research*, vol. 34, no. suppl 1, pp. D354–D357, 2006.
[28] X. Chen and L. Wang, "Integrating biological knowledge with gene expression profiles for survival prediction of cancer," *Journal of Computational Biology*, vol. 16, no. 2, pp. 265–278, 2009.
[29] N. Bandyopadhyay, T. Kahveci, S. Goodison et al., "Pathway-based feature selection algorithm for cancer microarray data," *Advances in Bioinformatics*, vol. 2009, Article ID 532989, 16 pages, 2009.
[30] M. L. Gatza, J. E. Lucas, W. T. Barry et al., "A pathway-based classification of human breast cancer," *Proceedings of the National Academy of Sciences*, vol. 107, no. 15, pp. 6994–6999, 2010.
[31] T. Y. Lin, "Granular computing on binary relation II: rough set representations and belief functions," *Rough Sets in Knowledge Discovery*, vol. 1, pp. 122–140, 1998.
[32] A. L. Prodromidis and S. J. Stolfo, "Cost complexity-based pruning of ensemble classifiers," *Knowledge and Information Systems*, vol. 3, no. 4, pp. 449–469, 2001.
[33] Z. H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: many could be better than all," *Artificial Intelligence*, vol. 137, no. 1, pp. 239–263, 2002.
[34] Z. Zhou and W. Tang, "Selective ensemble of decision trees," in *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*, pp. 476–483, Springer, Berlin, Germany, 2003.
[35] 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.
[36] E. K. Tang, P. N. Suganthan, and X. Yao, "An analysis of diversity measures," *Machine Learning*, vol. 65, no. 1, pp. 247–271, 2006.
[37] A. Bendavid, "Comparison of classification accuracy using Cohen's Weighted Kappa," *Expert Systems with Applications*, vol. 34, no. 2, pp. 825–832, 2008.
[38] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.
[39] G. Giacinto and F. Roli, "Design of effective neural network ensembles for image classification purposes," *Image and Vision Computing*, vol. 19, no. 9, pp. 699–707, 2001.
[40] A. Lazarevic and Z. Obradovic, "Effective pruning of neural network classifier ensembles," in *Proceedings of the IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)*, pp. 796–801, Washington, DC, USA, July 2001.
[41] S. Karim, *Exploring Plant Tolerance to Biotic and Abiotic Stresses*, Sveriges lantbruksuniv, Acta Universitatis agriculturae Sueciae, Uppsala, Sweden, 2007.
[42] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.