Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 972125, 8 pages
http://dx.doi.org/10.1155/2014/972125

Research Article

An Improved Feature Selection Based on Effective Range for Classification

Jianzhong Wang,[1,2] Shuang Zhou,[1,3] Yugen Yi,[1,4] and Jun Kong[1,3]

[1] *College of Computer Science and Information Technology, Northeast Normal University, Changchun 130000, China*
[2] *National Engineering Laboratory for Druggable Gene and Protein Screening, Northeast Normal University, Changchun 130000, China*
[3] *Key Laboratory of Intelligent Information Processing of Jilin Universities, Northeast Normal University, Changchun 130000, China*
[4] *School of Mathematics and Statistics, Northeast Normal University, Changchun 130000, China*

Correspondence should be addressed to Jun Kong; kongjun@nenu.edu.cn

Received 29 August 2013; Accepted 2 December 2013; Published 4 February 2014

Academic Editors: C.-C. Chang and J. Shu

Feature selection is a key issue in the domain of machine learning and related fields. The results of feature selection can directly affect the classifier's classification accuracy and generalization performance. Recently, a statistical feature selection method named effective range based gene selection (ERGS) is proposed. However, ERGS only considers the overlapping area (OA) among effective ranges of each class for every feature; it fails to handle the problem of the inclusion relation of effective ranges. In order to overcome this limitation, a novel efficient statistical feature selection approach called improved feature selection based on effective range (IFSER) is proposed in this paper. In IFSER, an including area (IA) is introduced to characterize the inclusion relation of effective ranges. Moreover, the samples' proportion for each feature of every class in both OA and IA is also taken into consideration. Therefore, IFSER outperforms the original ERGS and some other state-of-the-art algorithms. Experiments on several well-known databases are performed to demonstrate the effectiveness of the proposed method.

1. Introduction

Feature selection is widely used in the domain of pattern recognition, image processing, data mining, and machine learning before the tasks of clustering, classification, recognition, and mining [1]. In real-world applications, the huge dataset usually has a large number of features which contains much irrelevant or redundant information [1]. Redundant and irrelevant features cannot improve the learning accuracy and even deteriorate the performance of the learning models. Therefore, selecting an appropriate and small feature subset from the original features not only helps to overcome the "curse of dimensionality" but also contributes to accomplish the learning tasks effectively [2]. The aim of feature selection is to find a feature subset that has the most discriminative information from the original feature set. In general, feature selection methods are usually divided into three categories: embedded, wrapper, and filter methods [3, 4]. They are categorized based on whether or not they are combined with a specific learning algorithm.

In the embedded methods, the feature selection algorithm is always regarded as a component in the learning model. The most typical embedded based feature selection algorithms are decision tree approaches, such as ID3 [5], C4.5 [6], and CART algorithm [7]. In these algorithms, the features with the strongest ability of classification are selected in the nodes of the tree, and then the selected features are utilized to conduct a subspace to perform the learning tasks. Obviously the process of decision tree generation is also feature selection process.

Wrapper methods directly use the selected features to train a specific classifier and evaluate the selected subset according to the performance of the classifier. Therefore, the performances of wrapper methods strongly depend on the given classifier. Sequential forward selection (SFS) and sequential backward selection (SBS) [8] are two well-studied wrapper methods. SFS was initialized to an empty set. Then, the best feature from the complete feature set was chosen according to the evaluation criteria in each step and added into the candidate feature subset until it meets the stop condition. On the contrary, SBS started from the complete feature set. Then, it eliminated a feature which has the minimal impact on the classifier in each step until it satisfied the stop condition. Recently, Kabir et al. proposed a new wrapper based feature selection approach using neural network [9]. The algorithm was called constructive approach for feature selection (CAFS). The algorithm used a constructive approach involving correlation information to select the features and determine the architectures of neural network. Another wrapper based feature selection method was also proposed by Ye and Gong. In their approach, they considered the feature subset as the evaluation unit and the subset's convergence ability was utilized as the evaluation standard [10] for feature selection.

Different from the embedded and wrapper based algorithms, filter based feature selection methods directly select the best feature subset based on the intrinsic properties of the data. Therefore, the process of feature selection and learning model is independent in them. At present, the algorithms of filter based feature selection can be divided into two classes [11]: ranking and space searching. For the former, the feature selection process can be regarded as a ranking problem. More specifically, the weight (or score) of each feature is firstly computed. Then, the top k features are selected according to the ascending order of weight (or score). Pearson Correlation Coefficient (PCC) [12], Mutual Information (MI) [13], and Information Gain (IG) [14] are three commonly used ranking criterion to measure the dependency between each feature and the target variable. Another ranking criterion method named Relief [15], which analyzed the importance of each feature by computing the relationship between an instance and its nearest neighbors from the same and different classes, was proposed by Kira and Rendell. Then, an extension of Relief termed Relief-F was developed in [16]. Besides, there also exist many other methods proposed for ranking based filter feature selection. For more details about these algorithms, the readers can refer to [3, 4]. Although the ranking based filter methods have been applied to some real-world tasks successfully, a common shortcoming of these methods is that the feature subset selected by them may contain redundancy. In order to solve this problem, some space searching based filter methods have been proposed to remove the redundancy during feature selection. Correlation-based feature selection (CFS) [17] is a typical space searching algorithm; it did not only consider the correlation among features but also take the correlation between features and classes into account. Thus, CFS inclined to select the subset contains features that are highly correlated with the class and uncorrelated with each other. Minimum redundancy maximum relevance (MRMR) [18] is another method presented to reduce the redundancy of the selected feature subset.

Since both embedded and wrapper based feature selection methods interact with the classifier, they can only select the optimal subset for a particular classifier. So the features selected by them may be worse for other classifiers. Moreover, another disadvantage of the two methods is that they are more time consuming than filter method. Therefore, filter method is more fit for dealing with data that has large amounts of features since it has a good generalization ability [19]. As a result, we mainly focus on the research for filter based feature selection in this work.

In this paper, an integrated algorithm named Improved feature selection based on effective range (IFSER) is proposed for filter based feature selection. Our IFSER can be considered as an extension of the study in [20]. In [20], Chandra and Gupta presented a new statistical feature selection method named effective range based gene selection (ERGS). ERGS utilized the effective range of statistical inference theory [21] to calculate the weight of each feature, and a higher weight was assigned to the most important feature to distinguish different classes. However, since ERGS only considered the overlapping area (OA) among effective range of each class for every feature, it fails to handle the other relationships among the features of different classes. In order to overcome this limitation, the concept of including area (IA) is introduced into the proposed IFSER to characterize the inclusion relationship of effective ranges. Moreover, the samples' proportion for each feature of every class in both OA and IA is also taken into consideration in our IFSER. Therefore, IFSER outperforms the original ERGS and some other state-of-the-art algorithms. Experiments on several well-known databases are performed to demonstrate the effectiveness of the proposed method.

The rest of this paper is organized as follows. Section 2 briefly reviews ERGS and effective range. The proposed IFSER is introduced in Section 3. Section 4 reports experimental results on four datasets. Finally, we provide some conclusions in Section 5.

2. A Briefly Review on ERGS

In this section, we will review the effective range and ERGS algorithm briefly [20].

Let $F = \{F_i\}$ be the feature set of the dataset $X \in R^{N\times d}$, $i = 1, 2, \ldots, d$. $Y = \{Y_j\}$ $(j = 1, 2, \ldots, l)$ is the class labels of X. The class probability of jth class Y_j is p_j. For each class Y_j of the ith feature F_i, μ_{ij} and σ_{ij} denote the mean and standard deviation of the ith feature F_i for class Y_j, respectively. Effective range (R_{ij}) of jth class Y_j for ith feature F_i is defined by

$$R_{ij} = \left[r_{ij}^{-}, r_{ij}^{+}\right] = \left[\mu_{ij} - \left(1 - p_j\right)\gamma\sigma_{ij}, \mu_{ij} + \left(1 - p_j\right)\gamma\sigma_{ij}\right], \tag{1}$$

where r_{ij}^{-} and r_{ij}^{+} are the lower and upper bounds of the effective range, respectively. The prior probability of jth class is p_j. Here, the factor $(1 - p_j)$ is taken to scale down effect of class with high probabilities and consequently large variance. The value of γ is determined statistically by Chebyshev inequality defined as

$$P\left(\left|X - \mu_{ij}\right| \geq \gamma\sigma_{ij}\right) \leq \frac{1}{\gamma^2} \tag{2}$$

which is true for all distributions. The value of γ is set as 1.732 for the effective range which contains at least 2/3rd of the data objects [20].

Overlapping area (OA_i) among classes of feature F_i is computed by

$$\mathrm{OA}_i = \sum_{j=1}^{l-1} \sum_{k=j+1}^{l} \varphi_i(j,k), \tag{3}$$

where $\varphi_i(j,k)$ can be defined as

$$\varphi_i(j,k) = \begin{cases} r_{ij}^{+} - r_{ik}^{-} & \text{if } r_{ij}^{+} > r_{ik}^{-} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

In ERGS, for a given feature, the effective range of every class is first calculated. Then, the overlapping area of the effective ranges is calculated according to (3), and the area coefficient is computed for each feature. Next, the normalized area coefficient is regarded as the weight for every feature and an appropriate number of features are selected on the basis of feature weight. For more detailed information about the ERGS algorithm, the readers can refer to [20].

3. Improved Feature Selection Based on Effective Range

In this section, we present our improved feature selection based on effective range (IFSER) algorithm, which integrates overlapping area, including area and the samples' proportion for each feature of every class, into a unified feature selection framework.

3.1. Motivation. Although ERGS considers the overlapping area of every class for each feature, it fails to handle the problem of the inclusion relation of effective ranges. The problem is very realistic in real-world applications. Taking the gene data set as an example, Figure 1 shows the effective ranges of two gene samples from the Leukemia2 [22] gene database. From this figure, we can see that the overlapping area of gene number 9241 in Figure 1(a) is 165.7, and the overlapping area of gene number 3689 in Figure 1(b) is 170.8. Since the two overlapping areas of these two genes are similar, their weights obtained by ERGS are also similar. However, the relationships between the effective ranges in these two genes are very different. In Figure 1(a), the effective range of class 1 is completely included in the effective range of class 2, while the effective range of class 1 is partly overlapping with the effective range of class 2 in Figure 1(b). Therefore, the weight of the gene number 9241 in Figure 1(a) should be less than that in Figure 1(b) since all the samples in class 1 cannot be corrected and classified in this case. For this reason, the inclusion relation between the effective ranges (including area) must be taken into consideration.

Another example is shown in Figure 2. As can be seen from this figure, it is clearly found that the two features in Figures 2(a) and 2(b) have the same size of the overlapping area. However, the number of samples in these two areas is very different. In Figure 2(a), the number of samples belonging to the overlapping area is small but the number of samples belonging to the overlapping area in Figure 2(b) is relatively large. Thus, it is obvious that feature 1 is more important than feature 2 since more samples can be correctly classified. In other words, the weight assigned to feature1 should be greater than that assigned to feature 2. From this example, we can see that the samples' proportion for each feature of every class in both overlapping and including areas is also a vital factor to influence the features' weights and should be considered in the feature selection process.

3.2. Improved Feature Selection Based on Effective Range. Similar to ERGS, we suppose $F = \{F_i\}$ is the feature set of the dataset $X \in R^{N\times d}$, $i = 1, 2, \ldots, d$. $Y = \{Y_j\}$ ($j = 1, 2, \ldots, l$) is the class label set of the data samples in X. The class probability of jth class Y_j is p_j. For each class Y_j of ith feature F_i, μ_{ij} and σ_{ij} denote the mean and standard deviation of the ith feature F_i in class Y_j, respectively.

The first step of our proposed IFSER is to calculate the effective range of every class by

$$R_{ij} = \left[r_{ij}^{-}, r_{ij}^{+}\right] = \left[\mu_{ij} - \left(1 - p_j\right)\gamma\sigma_{ij}, \mu_{ij} + \left(1 - p_j\right)\gamma\sigma_{ij}\right], \tag{5}$$

where the definitions of r_{ij}^{-}, r_{ij}^{+}, p_j, and $1 - p_j$ are the same as those in ERGS.

The second step of our IFSER is to calculate overlapping areas OA_i among classes of feature F_i ($i = 1, 2, \ldots, d$) by

$$\mathrm{OA}_i = \sum_{j=1}^{l-1} \sum_{k=j+1}^{l} \varphi_i(j,k), \tag{6}$$

where the definition of $\varphi_i(j,k)$ is as same as in ERGS.

The third step of our proposed IFSER is to compute including area IA_i among classes of feature F_i ($i = 1, 2, \ldots, d$) by

$$\mathrm{IA}_i = \sum_{j=1}^{l-1} \sum_{k=j+1}^{l} \psi_i(j,k), \tag{7}$$

where $\psi_i(j,k)$ can be defined as

$$\psi_i(j,k) = \begin{cases} r_{ik}^{+} - r_{ik}^{-} & \text{if } r_{ij}^{+} \geq r_{ik}^{+} \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

The fourth step of our proposed IFSER is to compute area coefficient (AC_i) of feature F_i ($i = 1, 2, \ldots, d$) as

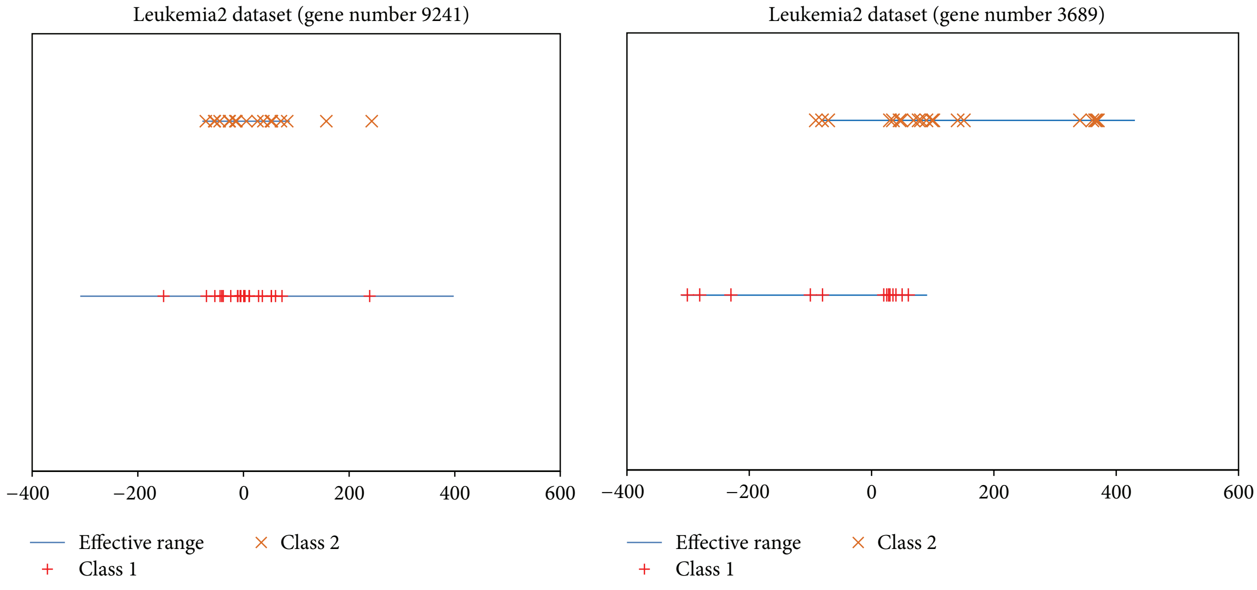
$$\mathrm{AC}_i = \frac{\mathrm{SA}_i}{\mathrm{Max}_j\left(r_{ij}^{+}\right) - \mathrm{Min}_j\left(r_{ij}^{-}\right)}, \tag{9}$$

where $\mathrm{SA}_i = \mathrm{OA}_i + \mathrm{IA}_i$. Then, the normalized area coefficient (NAC_i) can be obtained by

$$\mathrm{NAC}_i = 1 - \frac{\mathrm{AC}_i}{\max\left(\mathrm{AC}_s\right)}, \quad \text{for } s = 1, 2, \ldots, d. \tag{10}$$

From (10), we can clearly see that the features with larger NAC values are more important for distinguishing different classes.

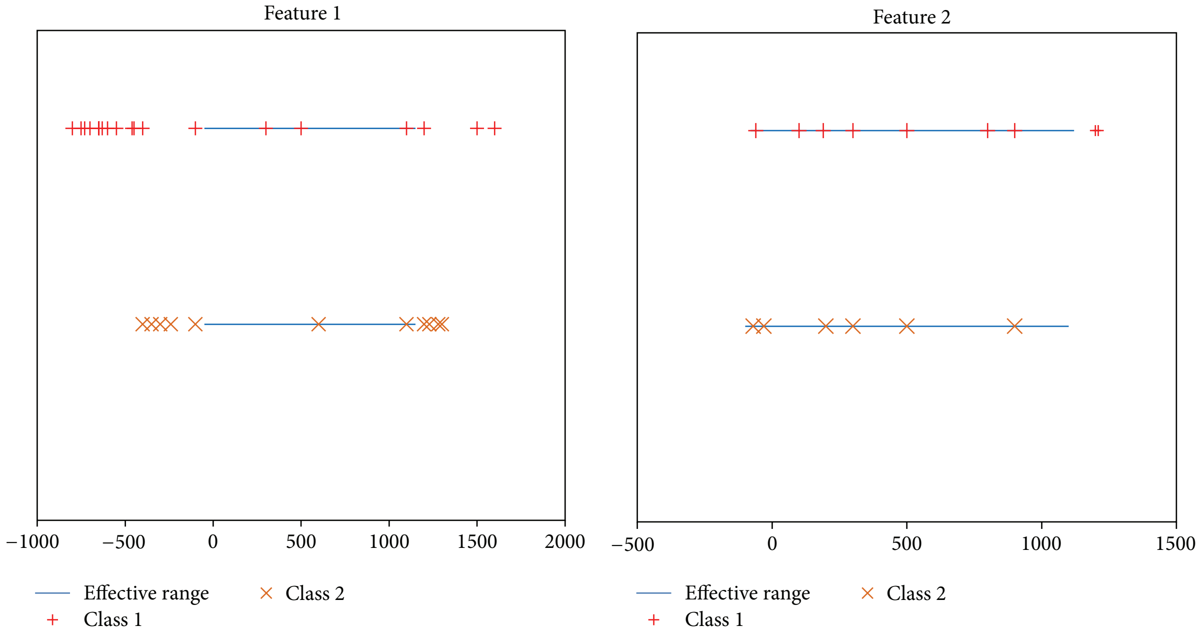
The fifth step of our proposed IFSER is to calculate the samples' number of each class in OA_i and IA_i for each feature



(a)

(b)

Figure 1: The ER of the gene accessions numbers 9241 and 3689 from the Leukemia2 gene database.



(a)

(b)

Figure 2: Different feature with the same size of overlapping area but different sample proportions in the two areas.

F_i. Let H_{ij} and G_{ij} denote samples' numbers of the jth class in OA_i and IA_i for each feature F_i. Assume that K_j represents the number of samples in the jth class. Then we use K_j divided by H_{ij} and G_{ij} to represent the proportions of samples in OA_i and IA_i, and for all classes of each feature the sums of the H_{ij}/K_j and G_{ij}/K_j are written as H_i and G_i.

For all classes of each feature F_i, the normalized H_i and G_i can be obtained by

$$\begin{aligned} NH_i &= 1 - \frac{H_i}{\max\left(H_s\right)} \\ GH_i &= 1 - \frac{G_i}{\max\left(G_s\right)}, \end{aligned} \quad \text{for } s = 1, 2, \ldots, d. \qquad (11)$$

From (11), the larger the value of NH_i and GH_i, the more significant the feature is.

The last step of our proposed IFSER is to compute the weight of each feature as

$$W_i = V_i \times Z_i, \tag{12}$$

where $V_i = \mathrm{NAC}_i$ and $Z_i = NH_i + GH_i$. From (12), we can see that a larger value of W_i indicates that the ith feature is more important. Therefore, we can select the features according to their weights and choose features with larger weights to form the selected feature subset.

Finally, the proposed IFSER algorithm can be summarized as in Algorithm 1.

4. Experiment and Results

In this section, in order to verify the performance of our proposed method, we conducted experiments on four datasets (Lymphoma [23], Leukemia1 [24], Leukemia2 [22], and 9_Tumors [25]) and compare our algorithm with five popular feature selection algorithms including ERGS [20], PCC [12], Relief-F [16], MRMR [18], and Information Gain [14]. Three classifiers are used to verify the effectiveness of our proposed method. The classification accuracies are obtained through leave-one-out cross-validation (LOOCV) in this work.

4.1. The Description of Datasets

4.1.1. Lymphoma Database. The Lymphoma database [23] consists of 96 samples and 4026 genes. There are two classes of samples in the dataset. The dataset comes from a study on diffuse large B-cell lymphoma.

4.1.2. Leukemia1 Database. Leukemia1 database [24] contains three types of Leukemia samples. The database has been constructed from 72 people who have acute myelogenous leukemia (AML), acute lymphoblastic leukemia (ALL) B cell, or ALL T-cell, and each sample is composed of 5327 gene expression profiles.

4.1.3. Leukemia2 Database. The Leukemia2 dataset [22] contains a total of 72 samples in three classes: AML, ALL, and mixed-lineage leukemia (MLL). The number of genes is 11225.

4.1.4. 9_Tumors Database. 9_Tumors database [25] consists of 60 samples of 5726 genes and categorized into 9 various human tumor types.

4.2. Experimental Results Using C4.5 Classifier. In this subsection, we estimate the performance of our proposed IFSER using C4.5 classifier on the four gene databases. Tables 1, 2, 3, and 4 summarize the results of the classification accuracies achieved by our methods and other methods. As we can see from Tables 1–4, the proposed IFSER method performs better than the other five algorithms in most cases. In particular, our proposed IFSER is much better than ERGS. The reason is that our proposed IFSER not only considers the overlapping area (OA) but also takes the including area and samples' proportion into account. These results demonstrate the fact that IFSER is able to select the best informative genes compared to other well-known techniques.

```
**Input**: Data matrix $X \in R^{N\times d}$, $i = 1, 2, \ldots, d$, the number of
selected feature $k$.
**Output**: Feature subset.
(1) Compute the ER of each feature by (5);
(2) Compute the $OA_i$ and $IA_i$ by (6) and (7);
(3) Compute the $AC_i$ by (9);
(4) Normalize the $AC_i$ by (10);
(5) Calculate the $NH_i$ and $GH_i$ by (11);
(6) Compute the weight of each feature by (12);
(7) Sort the weight of all features in a descending order;
(8) Select the best $k$ features;
```

Algorithm 1

For Lymphoma database, the classification accuracy of our proposed IFSER is substantial improvement compared with other algorithms. What is more, it is worth mentioning that our method only uses 10 features to achieve 93.75% classification accuracy. With the increase in feature dimension, the classification results of most methods (such as our proposed IFSER, PCC, IG, and ERGS) are reduced. For Relief-F and MRMR, the classification results are very low when the feature dimension is equal to 10 at the beginning. Then, with the increase in feature dimension, the classification results are improved. When they achieve the best results, the classification results begin to decrease with the increase in the dimension again.

For Leukemia1 and Leukemia2 databases, the performance of our proposed IFSER is also better than ERGS and other methods. Our proposed IFSER can achieve the best results when the feature dimension is between 50 and 70. For Leukemia1 database, the performances of MRMR and ERGS keep stable on most dimensions. The trend of the classification results of PCC on Leukemia2 is inconsistent with those on Lymphoma database since it is almost monotonously decreased with the increase of feature dimension. And the other results are consistent with the experiments on Lymphoma database.

For 9_Tumors database, as we can see from Table 4, the performances of all the methods are very low due to the fact that database only contains 60 samples but 5726 genes. However, the performance of our proposed IFSER is much better than other algorithms. This result demonstrates the fact that our proposed IFSER is able to deal with the small sample size and high dimensions gene data.

4.3. Experimental Results Using NN Classifier. In this subsection, we evaluate the performance of our proposed IFSER using nearest neighbor (NN) classifier on the four gene databases. The results of the classification accuracies achieved by our proposed and other methods are listed in Tables 5, 6, 7, and 8. Comparing Tables 5–8 with Tables 1–4, we can see that the classification results of all the methods are improved.

Table 1: Classification accuracies (%) of different feature selection methods with C4.5 on Lymphoma database.

	10	30	50	70	90	110	130
PCC	84.38	82.29	80.21	78.13	79.17	79.17	80.21
Relief-F	69.79	72.92	72.92	75.00	68.75	66.67	82.29
IG	78.13	76.04	76.04	72.92	77.08	77.08	77.08
MRMR	71.88	79.17	79.17	80.21	81.25	80.21	79.17
ERGS	86.46	85.42	82.29	81.25	83.33	83.33	84.38
IFSER	93.75	86.46	83.33	83.33	83.33	80.21	79.17

Table 2: Classification accuracies (%) of different feature selection methods with C4.5 on Leukemia1 database.

	10	30	50	70	90	110	130
PCC	88.89	88.89	88.89	87.50	87.50	87.50	87.50
Relief-F	75.00	79.17	75.00	80.56	81.94	79.17	80.56
IG	80.56	84.72	84.72	84.72	84.72	84.72	84.72
MRMR	84.72	84.72	84.72	84.72	84.72	84.72	84.72
ERGS	88.89	88.89	88.89	88.89	88.89	88.89	88.89
IFSER	84.72	84.72	86.11	90.28	90.28	88.89	87.50

Table 3: Classification accuracies (%) of different feature selection methods with C4.5 on Leukemia2 database.

	10	30	50	70	90	110	130
PCC	80.56	83.33	87.50	87.50	87.50	86.11	86.11
Relief-F	77.78	75.00	84.72	86.11	80.56	77.78	76.39
IG	84.72	87.50	87.50	87.50	87.50	87.50	87.50
MRMR	84.72	88.89	88.89	88.89	88.89	88.89	88.89
ERGS	86.11	84.72	88.89	88.89	88.89	87.50	87.50
IFSER	79.17	88.89	90.28	88.89	88.89	87.50	88.89

Table 4: Classification accuracies (%) of different feature selection methods with C4.5 on 9_Tumors database.

	10	30	50	70	90	110	130
PCC	28.33	28.33	26.67	25.00	28.33	26.67	28.33
Relief-F	20.00	16.67	30.00	28.33	31.67	36.67	36.67
IG	38.33	38.33	41.67	40.00	40.00	40.00	38.33
MRMR	38.33	38.33	40.00	36.67	38.33	40.00	40.00
ERGS	28.33	28.33	23.33	25.00	23.33	21.67	26.67
IFSER	25.00	36.67	43.33	48.33	46.67	43.33	43.33

Table 5: Classification accuracies (%) of different feature selection methods with NN on Lymphoma database.

	10	30	50	70	90	110	130
PCC	89.58	96.88	94.79	95.83	97.92	97.92	96.88
Relief-F	68.75	84.38	86.46	88.54	87.50	85.42	88.54
IG	88.54	95.83	94.79	94.79	95.83	96.88	96.88
MRMR	88.54	91.67	93.75	93.75	93.75	93.75	93.75
ERGS	89.58	94.79	95.83	97.92	95.83	97.92	97.92
IFSER	94.79	94.79	96.88	96.88	97.92	97.92	97.92

Table 6: Classification accuracies (%) of different feature selection methods with NN on Leukemia1 database.

	10	30	50	70	90	110	130
PCC	93.06	94.44	95.83	97.22	95.83	97.22	95.83
Relief-F	69.44	76.31	75.00	75.00	73.61	76.39	80.56
IG	93.06	94.44	91.67	93.06	93.06	94.44	93.06
MRMR	88.89	93.06	90.28	93.06	93.06	94.44	93.06
ERGS	94.44	95.83	94.44	95.83	95.83	95.83	95.83
IFSER	81.94	91.67	93.06	91.67	97.22	94.44	95.83

Table 7: Classification accuracies (%) of different feature selection methods with NN on Leukemia2 database.

	10	30	50	70	90	110	130
PCC	88.89	88.89	90.28	93.06	91.67	91.67	91.67
Relief-F	69.44	83.33	83.33	83.33	87.50	93.06	94.44
IG	83.33	83.33	94.44	94.44	94.44	94.44	94.44
MRMR	88.89	90.28	93.06	93.06	93.06	93.06	93.06
ERGS	86.11	86.11	93.06	93.06	91.67	93.06	93.06
IFSER	84.27	91.67	93.06	91.67	88.89	90.28	94.44

Table 8: Classification accuracies (%) of different feature selection methods with NN on 9_Tumors database.

	10	30	50	70	90	110	130
PCC	28.33	41.67	51.67	51.67	51.67	50.00	51.67
Relief-F	25.00	28.33	21.67	26.67	30.00	35.00	33.33
IG	48.33	51.67	60.00	58.33	60.00	61.67	58.33
MRMR	38.33	46.67	56.67	55.00	60.00	65.00	61.67
ERGS	25.00	30.00	40.00	38.33	41.67	41.67	45.00
IFSER	35.00	36.67	38.33	46.67	46.67	45.00	46.67

For Lymphoma database, IFSER, PCC, and ERGS are better than Relief-F, IG, and MRMR. For Leukemia1 database, our proposed IFSER and PCC outperform Relief-F, IG, MRMR, and ERGS. And the best result of IFSER is the same as PCC. However, for Leukemia2, IFSER, IG, and Relief-F achieve the best results than PCC, MRMR, and ERGS. For 9_Tumors database, the performance of IFSER is worse than PCC, IG, and MRMR, but better than Relief-F and ERGS. These results demonstrate the fact that result of feature selection depends on the classifier, and it is crucial to choose an appropriate classifier for different feature selection methods.

4.4. Experimental Results Using SVM Classifier. The performance of our proposed IFSER using support vector machine (SVM) classifier on the four gene database is tested in this subsection. Figures 3–6 show the classification accuracies of different algorithms on four gene databases. From Figures 3 and 4, we can see that our proposed IFSER outperforms other algorithms in most cases. And the IFSER achieves its best result at a lower dimension than other algorithms. This result further demonstrates the fact that IFSER is able to select the best informative genes as compared to other

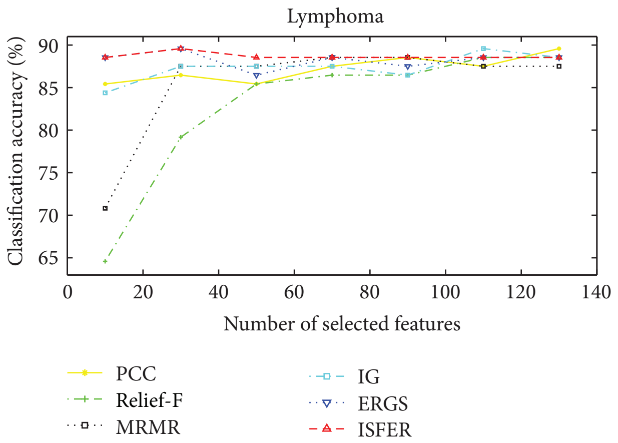


FIGURE 3: The classification accuracies of different algorithms on the Lymphoma database.

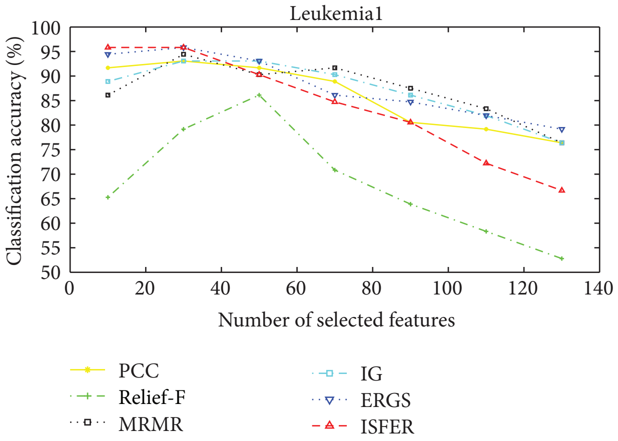


FIGURE 4: The classification accuracies of different algorithms on the Leukemia1 database.

feature selection techniques. As we can see from Figure 5, our proposed IFSER is worse than Relief-F, IG, MRMR and ERGS. From Figure 6, it is found that our proposed IFSRE outperforms PPC, Relief-F, IG, and ERGS but is not as good as MRMR. This indicates that the SVM classifier is not suitable for the feature selected results of our proposed algorithm on small sample size databases.

5. Conclusions

In this paper, we propose a novel statistical feature selection algorithm named effective range based gene selection (IFSER). Compared with existing algorithms, IFSER not only considers the overlapping areas of the features in different classes but also takes the including areas and the samples' proportion in overlapping and including areas into account. Therefore, IFSER outperforms the original ERGS and some other state-of-the-art algorithms. Experiments on several well-known databases are performed to demonstrate the effectiveness of the proposed method.

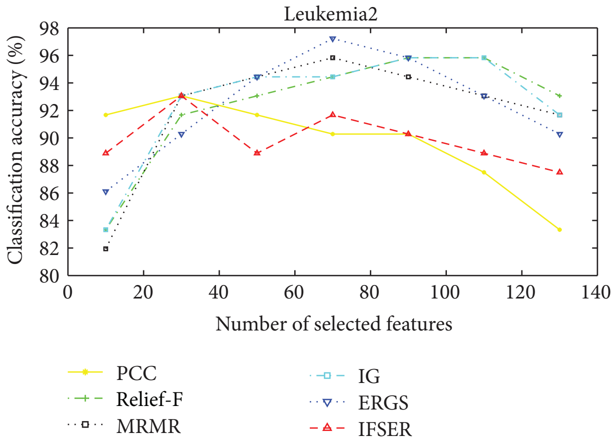


FIGURE 5: The classification accuracies of different algorithms on the Leukemia2 database.

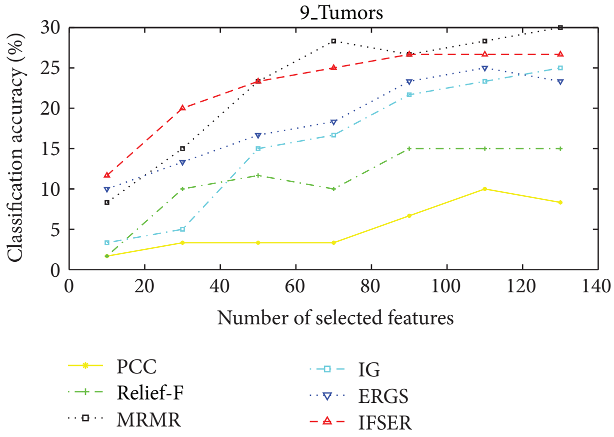


FIGURE 6: The classification accuracies of different algorithms on the 9_Tumors database.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by Fund of Jilin Provincial Science & Technology Department (nos. 201115003, 20130206042GX), Fundamental Research Funds for the Central Universities (no. 11QNJJ005), the Science Foundation for Postdoctor of Jilin Province (no. 2011274), and Young Scientific Research Fund of Jilin Province Science and Technology Development project (no. 20130522115JH).

References

[1] E. Xing, M. Jordan, and R. Karp, "Feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, pp. 1–12, 2005.

[2] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[3] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[4] C. Lazar, J. Taminau, S. Meganck et al., "A survey on filter techniques for feature selection in gene expression microarray analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 4, pp. 1106–1119, 2012.

[5] J. R. Quinlan, "Learning efficient classification procedures and their application to chess end games," in *Machine Learning: An Artificial Intelligence Approach*, pp. 463–482, Morgan Kaufmann, San Francisco, Calif, USA, 1983.

[6] J. R. Quinlan, *C4.5: Programs For Machine Learning*, Morgan Kaufmann, San Francisco, Calif, USA, 1993.

[7] L. Breiman, J. H. Friedman et al., *Classification and Regression Trees*, Wadsforth International Group, 1984.

[8] J. Kittler, "Feature set search algorithms," in *Pattern Recognition and Signal Processing*, C. H. Chen, Ed., pp. 41–60, Sijthoff and Noordhoff, The Netherlands, 1978.

[9] M. M. Kabir, M. M. Islam, and K. Murase, "A new wrapper feature selection approach using neural network," *Neurocomputing*, vol. 73, no. 16–18, pp. 3273–3283, 2010.

[10] J. X. Ye and X. L. Gong, "A novel fast Wrapper for feature subset selection," *Journal of Changsha University of Science and Technology*, vol. 7, no. 4, pp. 69–73, 2010.

[11] J. Wang, L. Wu, J. Kong, Y. Li, and B. Zhang, "Maximum weight and minimum redundancy: a novel framework for feature subset selection," *Pattern Recognition*, vol. 46, no. 6, pp. 1616–1627, 2013.

[12] L. J. Van't Veer, H. Dai, M. J. Van de Vijver et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[13] 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.

[14] H. Liu, J. Li, and L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns," *Genome Informatics Series*, vol. 13, pp. 51–60, 2002.

[15] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proceedings of the 9th International Conference on Machine Learning*, pp. 249–256, 1992.

[16] I. Kononenko, "Estimating features: analysis and extension of RELIEF," in *Proceedings of the 6th European Conference on Machine Learning*, pp. 171–182, 1994.

[17] M. A. Hall, *Correlation-based feature selection for machine learning [Ph.D. thesis]*, University of Waikato, Hamilton, New Zealand, 1999.

[18] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 2, pp. 185–205, 2005.

[19] H. Almuallim and T. Dietterich, "Learning with many irrelevant features," in *Proceedings of the 9th National Conference on Artificial Intelligence*, pp. 547–552, San Jose, 1991.

[20] B. Chandra and M. Gupta, "An efficient statistical feature selection approach for classification of gene expression data," *Journal of Biomedical Informatics*, vol. 44, no. 4, pp. 529–535, 2011.

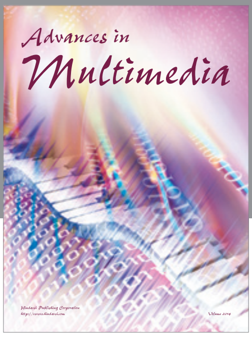
[21] W. Härdle and L. Simar, *Applied Multivariate Statistical Analysis*, Springer, 2007.

[22] S. A. Armstrong, J. E. Staunton, and L. B. Silverman, "*MLL* translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature Genetics*, vol. 30, no. 1, pp. 41–47, 2002.

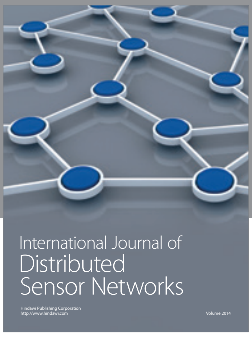
[23] A. Alizadeh et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503–511, 2000.

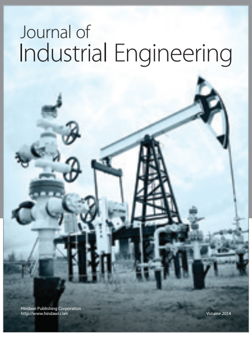
[24] Chemosensitivity prediction by transcriptional profiling, *Whitehead Massachusetts Institute of Technology Center For Genome Research*, Cambridge, Mass, USA.

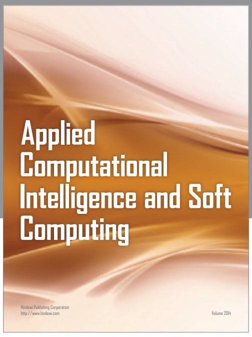
[25] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–527, 1999.

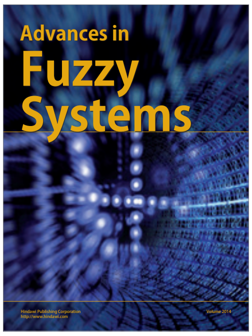

Advances in
Multimedia

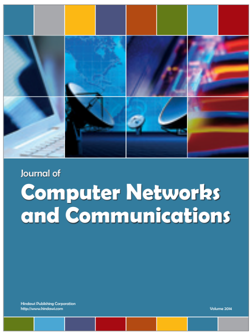
The Scientific
World Journal

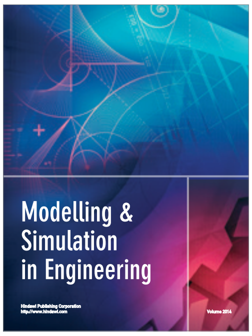

International Journal of
Distributed
Sensor Networks

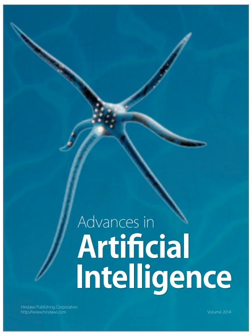

Journal of
Industrial Engineering

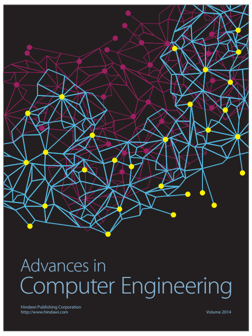

Applied
Computational
Intelligence and Soft
Computing


Advances in
Fuzzy
Systems

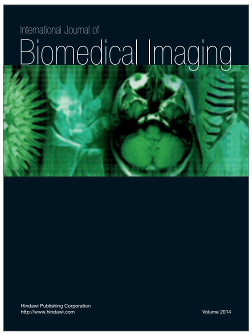

Journal of
Computer Networks
and Communications

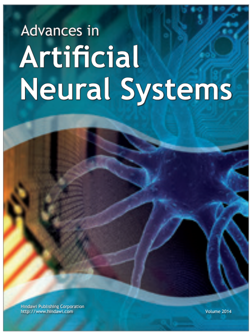

Modelling &
Simulation
in Engineering

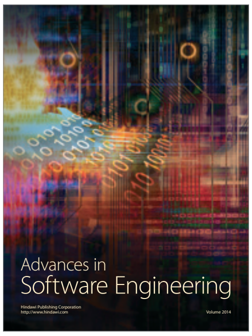

Advances in
Artificial
Intelligence


Advances in
Computer Engineering

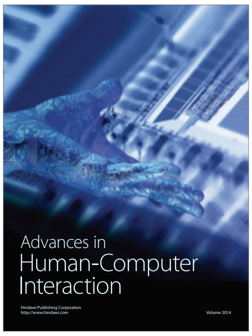
International Journal of
Computer Games
Technology

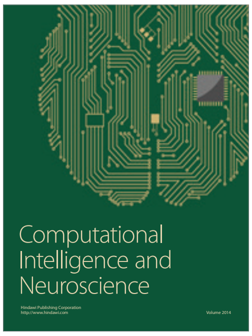

International Journal of
Biomedical Imaging


Advances in
Artificial
Neural Systems


Advances in
Software Engineering

Journal of
Robotics


Advances in
Human-Computer
Interaction


Computational
Intelligence and
Neuroscience

International Journal of
Reconfigurable
Computing

Journal of
Electrical and Computer
Engineering