International Journal of Hybrid Information Technology
Vol.8, No.7 (2015), pp.277-292
http://dx.doi.org/10.14257/ijhit.2015.8.7.26

A Novel Feature Gene Selection Method Based On Neighborhood Mutual Information

Tao Chen[1, 2*],Zenglin Hong[1],Hui Zhao[2],Xiao Yang[2], Jun Wei[2]

1 School of Automation, Northwestern Polytechnical University, Xi'an, Shaanxi, 710072, China
2 School of Mathematics and Computer Science, Shaanxi University of Technology, Hanzhong, Shaanxi, 723000, China
**Corresponding author:ct79hz@126.com*

Abstract

DNA microarray technique can detect tens of thousands of genes activity in cells and has been widely used in clinical diagnosis. However, microarray data has characteristics of high dimension and small samples, moreover many irrelevant and redundant genes also decrease performance of classification algorithm .Mutual information is very effective method and has widely been used in feature gene selection, but it cannot directly deal with continuous features. Therefore, this paper proposes a novel feature gene selection method to resolve this problem. Firstly, a lot of irrelevant genes are eliminated from original data by using reliefF algorithm , and the candidate subset of genes is obtained; Secondly, a algorithm based on neighborhood mutual information and forward greedy search strategy which deals with directly continuous features is proposed to select feature genes in above genes subset. Here, because radius of neighborhood greatly affects reduction performance, differential evolution algorithm is applied to optimize radius before reduction. The simulation results on six benchmark microarray datasets show that our method can obtain higher classification accuracy using as few genes as possible, especially neighborhood mutual information can directly continuous features. Feature genes selected has an important meaning for understanding microarray data and finding pathogenic genes of cancer. It is an effective and efficient method for feature genes selection.

Keywords: *feature gene selection; ReliefF algorithm; neighborhood mutual information; differential evolution algorithm*

1. Introduction

DNA microarray is a high-throughput technology and is used to measure the expression levels of thousands of genes simultaneously. The fundamental idea is to exploit complementary base pairing to measure the amount of the different types of mRNA molecules in a cell, thus indirectly measuring the expression levels of the genes that are responsible for the synthesis of those particular mRNA molecules [1, 2].DNA microarray technology can help researchers to learn more about many different diseases related to genes, including heart disease, mental illness, especially cancer. In the past, scientists have classified different types of cancers based on the organs. However, with the development of microarray technology, they are able to further classify these types of cancers based on the patterns of gene activity in the tumor cells. Therefore, microarray technology is important meaning for the diagnosis and pathogenic mechanism of disease related to genes.

Microarray data often contains a small number of samples (tens or hundreds) and a large number of genes (tens of thousands), which leads to imbalance between the number of samples and genes, and causes the curse of dimensionality [3, 4]. Many researches show the classification performance reaches the highest level when the dimensionality increases to a certain number, and then the classification performance decreases slowly with the increase of dimensionality. The reason is the number of samples for learning is limited. Therefore, in order to ensure the classification performance, it needs to increase the number of samples. But the number of samples needed in the learning process increase by the exponential growth mode with the increase of the number of features. It is difficult to obtain large number of samples because of the high cost of microarray experiments. In addition, the microarray data contains many irrelevant and redundant genes, which not only increase the dimension of feature space and decrease the learning efficiency, but also increase the possibility of noise data, and it interferes with the learning process of classification algorithm and affects the construction and the results of classification model. Therefore, in order to decrease the influence of adverse factors and reduce the dimension of feature space, the irrelevant and redundant genes should be removed from original data. It can effectively improve the efficiency and performance of algorithm to avoid over fitting phenomena, and gene selection is an effective method to solve the above problems.

Gene selection refers to the process of removing irrelevant and redundant genes and preserving those feature genes to classify microarray data [5]. The most commonly gene selection approaches are based on ranking. Each gene is evaluated individually and assigned a score to reflect its correlation with the class according to certain criteria. And then genes are ranked according to their scores and the top-ranked genes are selected as feature genes. Feature gene selection methods ranking-based mainly include t statistic[6], information gain[7,8], χ^2 statistic[9], the threshold number of misclassification score [10], concatenation of several feature filtering [11], Rrelief [12], ReliefF [13], mutual information [14-16],entropy [17] and correlation based feature selection (CFS) [18] ,etc.

Mutual information (MI) is widely used because of effectiveness of genes selection [15, 16]. MI is applied to measure the amount of information that one random variable contains about another random variable and reflects the degree of linear or nonlinear dependency between variables. In the process of computing MI, we need to know probability distributions of variables and their joint distribution. However, probability distributions are not known in practice. As to a set of samples, we have to estimate the probability distributions and joint distributions of features. If features are discrete, we use histogram to estimate the probability. The probabilities are computed as the relative frequency of samples with the corresponding feature values. If features are continuous, there are two methods to estimate probability. One is Parzen Window to estimate information quantity [19], but it is usually difficult to obtain accurate estimates for multivariate density samples in high-dimensional space because of sparse distribution. Moreover the computational cost is also very high [20].Genes data in microarray are continuous and sparse,and it leads to limitation of Parzen Window for dealing with genes data. The other method is discretization, and the domains of variables are partition into several subsets by discretization technology[21].But shen pointed that discretization process may lead to some information loss from the original data and it affects classification accuracy, and the reason is information is lost and not fully utilized[22].

For above problems, we introduce a new measure of relevance between continuous genes and discrete decision features, called neighborhood mutual information (NMI) [23, 24]. NMI is constructed by integrating the concept of neighborhood into Shannon’s information theory and is a natural generalization of mutual information in numerical feature spaces.

This paper proposes a novel feature gene selection method, which includes two stages. The first stage is genes preselecting. Candidate gene subset which has high relevance with classification task is obtained by using reliefF. The second stage is feature genes selection. In this stage, because radius of neighborhood in neighborhood mutual information highly affects performance of algorithm, so differential evolution algorithm [25, 26] is used to optimize radius of neighborhood firstly, and then feature genes are selected by using optimized neighborhood mutual information.

The remainder of this paper is organized as follows: The reliefF is introduced in section 2. Neighborhood mutual information is defined, and then a feature selection algorithm NMI_FGS based on neighborhood mutual information and forward greedy search strategy is proposed in section 3.Section 4 applies differential evolution algorithm to optimize radius of neighborhood mutual information in NMI_FGS. Section 5 gives ideas and steps of our proposed method. Section 6 makes experiment on six benchmark microarray datasets and gives the experimental results and analysis. The conclusion is given in section 7.

2. ReliefF Algorithm

Relief algorithm [12] was proposed by Kira and Rendell in 1994.The success of the algorithm is due to the fact that it is fast, easy to understand and implement and accurate even with dependent features and noisy data. But relief can only be applied to two classification problems, so Kononenko proposed reliefF [13], which is the extension of relief. The key idea of reliefF is to estimate the quality of attributes according to how well their values distinguish between the instances that are near to each other. For that purpose, given a randomly selected instance x ,ReliefF searches for k nearest neighbors of x from the same class, called nearHist, and also k nearest neighbors of x from each of the different class, called nearMisses. The quality estimation $W(g)$ for each attribute g is update formula, nearHist and nearMisses. In the update formula, the contributions of all the hits and misses are averaged.The process is repeated for n times to return weights of all features. ReliefF is fast, not limited by data types, fairly noise-tolerant, and unaffected by feature interaction, but it does not deal with redundant feature.

3. Neighborhood Mutual Information

3.1 Mutual Information

Mutual information is widely used in quantifying linear or nonlinear relevance degree between random variables.

For discrete random variables, we give some definitions about entropy and mutual information [15, 23, and 24].

Definition 1 $A=\{a_1,a_2,\dots a_n\}$ is a discrete random variable, $p(a_i)$ is the probability of a_i ,the entropy of A is defined as $H(A)=-\sum_{i=1}^{n} p(a_i)\log p(a_i)$.

Definition 2 $A=\{a_1,a_2,\dots a_n\}$ and $B=\{b_1,b_2,\dots b_m\}$ are two discrete random variables, $p(a_i,b_j)$ is the joint probability of a_i and b_j , $i=1,2,\dots,n; j=1,2,\dots,m$ .The joint entropy of A and B is defined as $H(A,B)=-\sum_{i=1}^{n}\sum_{j=1}^{m} p(a_i,b_j)\log p(a_i,b_j)$.

Definition 3 $A=\{a_1,a_2,\dots a_n\}$ and $B=\{b_1,b_2,\dots b_m\}$ are two discrete random variables. If B is known, the conditional entropy of A conditioned to B , that is the uncertainty of A , is defined as $H(A|B)=H(A,B)-H(B)=-\sum_{i=1}^{n}\sum_{j=1}^{m} p(a_i,b_j)\log p(a_i|b_j)$.

Definition 4 $A=\{a_1,a_2,\dots a_n\}$ and $B=\{b_1,b_2,\dots b_m\}$ are two discrete random variables. The mutual information between A and B , that is the reduction of uncertainty of A due to the information of B , is defined as $MI(A;B)=\sum_{i=1}^{n}\sum_{j=1}^{m} p(a_i,b_j)\log\frac{p(a_i|b_j)}{p(a_i)}$.

Theorem 1 $MI(A;B)=MI(B;A)=H(A)+H(B)-H(A,B)$

Proof. As $\frac{p(a_i|b_j)}{p(a_i)}=\frac{p(b_j|a_i)}{p(b_j)}=\frac{p(a_i,b_j)}{p(a_i)p(b_j)}$, we have

$$\begin{aligned}MI(A;B)&=MI(B;A)=H(A)-H(A|B)=H(B)-H(B|A)\\&=H(A)+H(B)-H(A,B)\end{aligned}.$$

For continuous random variables, corresponding concepts is as follows [21].

Definition 5 A and B are two continuous random variables, $p(a)$ and $p(b)$ are the probability density function of a, b respectively. $p(a,b)$ is the joint probability density function of a and b. The entropy of A is defined as $H(A)=-\int p(a)\log p(a)da$.

Mutual information between A and B is defined as $MI(A;B)=\iint p(a)\log\frac{p(a,b)}{p(a)p(b)}dadb$.

3.2 Neighborhood Mutual Information

It is difficult to estimate probability density of continuous features in the process of calculating mutual information. The idea of neighborhood is introduced into Shannon's entropy, and neighborhood mutual information is defined [23, 24]. And on this basis, this paper proposes a feature selection algorithm based on neighborhood mutual information and forward greedy search strategy.

Definition 6 $U=\{x_1,x_2,...,x_n\}$, $x_i\in R^N$ is a samples set. The neighborhood of sample x is defined as $\delta(x)=\{x_i|\Delta(x,x_i)\le\delta\}$, where δ is a constant and $\delta\ge 0$. Δ is a distance function. $\forall x_1,x_2,x_3\in U$, it satisfies:

(1) $\Delta(x_1,x_2)\ge 0$, $\Delta(x_1,x_2)=0$ if and only if $x_1=x_2$;

(2) $\Delta(x_1,x_2)=\Delta(x_2,x_1)$; (3) $\Delta(x_1,x_2)+\Delta(x_2,x_3)\ge\Delta(x_1,x_3)$.

Definition 7 $U=\{x_1,x_2,...,x_n\}$ is a samples set described by features set $F=\{f_1,f_2,\dots f_m\}$. $S\subseteq F$ is a subset of features. The neighborhood of sample x_i in S is denoted by $\delta_S(x_i)=\{x|\Delta(x,x_i)\le\delta,x\in U_S\}$.

The neighborhood uncertainty of x_i is defined as $NH_\delta^{x_i}(S)=-\log\frac{\|\delta_s(x_i)\|}{n}$.

The average uncertainty of the samples set U is defined as $NH_\delta(S)=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_s(x_i)\|}{n}$.

Where $\|X\|$ is the cardinality of the set X.

As for $\forall x_i,\delta_S^{x_i}\subseteq U,\frac{\|\delta_s(x_i)\|}{n}\le 1$, so we have $\log n\ge NH_\delta(S)\ge 0$. In addition, if $\delta_1\le\delta_2$, we have $\delta_1(x_i)\subseteq\delta_2(x_i)$. So $NH_{\delta_1}^{x_i}(S)\ge NH_{\delta_2}^{x_i}(S)$ and $NH_{\delta 1}(S)\ge NH_{\delta 2}(S)$.

In addition, It is easy to obtain that $NH_\delta(S)=\log n$ if and only if for $\forall x_i$, $\|\delta_S(x_i)\|=1$;

$NH_\delta(S)=0$ if and only if for $\forall x_i$, $\|\delta_S(x_i)\|=n$.

Theorem 2 If $\delta_1=0$, then $NH_\delta(S)=H(S)$, where $H(S)$ is shannon's entropy.

Proof. If $\delta_1=0$, the samples are divided into disjoint $X_1, X_2,\mathrm{K}\,X_m$. Where $\forall x_i,x_j\in X_k$, $\Delta(x_i,x_j)=0$.

Assumed there are m_i samples in X_i, then if $\forall x\in X_k$ and $\delta=0$, $H(S)=-\sum_{i=1}^{m}\frac{m_i}{n}\log\frac{m_i}{n}$,

$\delta_S(x) = X_k$ .If $i \neq j$, $X_i \cap X_i = \varnothing$,we have $NH_\delta(S) = -\frac{1}{n}\sum_{i=1}^{m}\log\frac{\|\delta_s(x_i)\|}{n} = -\frac{\|X_j\|}{n}\sum_j \log\frac{\|X_j\|}{n} = H(S)$

Definition 8 $U = \{x_1, x_2, ..., x_n\}$ is a samples set, and $F = \{f_1, f_2, \dots f_m\}$ is a features set. $R, S \subseteq F$ are two subsets of features. The neighborhood of sample x_i in feature subspace $R \cup S$ is denoted by $\delta_{R\cup S}(x_i)$,then the joint neighborhood entropy of $R \cup S$ is defined as

$$NH_\delta(R,S) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_{R\cup S}(x_i)\|}{n} .$$

Especially, if R is a set of input variables and C is the decision attribute. we define $\delta_{R\cup C}(x_i) = \delta_R(x_i) \cap C(x_i)$,then $NH_\delta(R,C) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_R(x_i) \cap C(x_i)\|}{n}$.

Theorem 3 $NH_\delta(R,S) \geq NH_\delta(R), NH_\delta(R,S) \geq NH_\delta(S)$.

Proof. $\forall x \in U$,we have $\delta_{S\cup R}(x_i) \subseteq \delta_S(x_i)$ and $\delta_{S\cup R}(x_i) \subseteq \delta_R(x_i)$ .Then $\|\delta_{S\cup R}(x_i)\| \leq \|\delta_S(x_i)\|$

and $\|\delta_{S\cup R}(x_i)\| \leq \|\delta_R(x_i)\|$ .Therefore $NH_\delta(R,S) \geq NH_\delta(R), NH_\delta(R,S) \geq NH_\delta(S)$.

Definition 9 $U = \{x_1, x_2, ..., x_n\}$ is a samples set and $F = \{f_1, f_2, \dots f_m\}$ is a features set. $R, S \subseteq F$ Are two subsets of features. The conditional neighborhood entropy of R to S is defined as

$$NH_\delta(R|S) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_{S\cup R}(x_i)\|}{\|\delta_S(x_i)\|} .$$

Theorem 4 $NH_\delta(R|S) = NH_\delta(R,S) - NH_\delta(S)$.

Proof. $NH_\delta(R,S) - NH_\delta(S)$

$$\begin{aligned}
&= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_{S\cup R}(x_i)\|}{n} - (-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_S(x_i)\|}{n}) \\
&= -\frac{1}{n}\sum_{i=1}^{n}(\log\frac{\|\delta_{S\cup R}(x_i)\|}{n} - \log\frac{\|\delta_S(x_i)\|}{n}) \\
&= -\frac{1}{n}\sum_{i=1}^{n}(\log\frac{\|\delta_{S\cup R}(x_i)\|}{\|\delta_S(x_i)\|}) \\
&= NH_\delta(R|S)
\end{aligned}$$

Definition 10 $U = \{x_1, x_2, ..., x_n\}$ is a samples set and $F = \{f_1, f_2, \dots f_m\}$ is a features set. $R, S \subseteq F$

are two subsets of features. The neighborhood mutual information of R and S is defined as

$$NMI_\delta(R;S) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_R(x_i)\| \cdot \|\delta_S(x_i)\|}{n\|\delta_{S\cup R}(x_i)\|} .$$

Theorem 5 R and S are two subsets of features, and $NMI_\delta(R;S)$ is the neighborhood mutual information of R and S .The following equations hold:

(1) $NMI_\delta(R;S) = NMI_\delta(S;R)$;

(2) $NMI_\delta(R;S) = NH_\delta(R) + NH_\delta(S) - NH_\delta(R,S)$;

(3) $NMI_\delta(R;S) = NH_\delta(R) - NH_\delta(R|S) = NH_\delta(S) - NH_\delta(S|R)$

Proof. The equation (1) and (3) are straightforward. we give the proof of equation (2).

$$NH_\delta(R) + NH_\delta(S) - NH_\delta(R,S)$$

$$=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_R(x_i)\|}{n}-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_S(x_i)\|}{n}-(-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_{R\cup S}(x_i)\|}{n})$$

$$=-\frac{1}{n}\sum_{i=1}^{n}(\log\frac{\|\delta_R(x_i)\|}{n}+\log\frac{\|\delta_S(x_i)\|}{n}-\log\frac{\|\delta_{R\cup S}(x_i)\|}{n})$$

$$=-\frac{1}{n}\sum_{i=1}^{n}(\log\frac{\|\delta_R(x_i)\|}{n}\cdot\frac{\|\delta_S(x_i)\|}{n}\cdot\frac{n}{\|\delta_{R\cup S}(x_i)\|})$$

$$=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_R(x_i)\|\cdot\|\delta_S(x_i)\|}{n\|\delta_{R\cup S}(x_i)\|}$$

Lemma 1 U is a samples set described by the features F . $R\subseteq F$ is a subset of features and C is the decision attribute. $NMI_{\delta}^{x}(R;C)=H^{x}(C)$ if the decision of sample $x\in U$ is δ - neighborhood consistent, where

$$NMI_{\delta}^{x}(R;C)=-\log\frac{\|\delta_R(x)\|\cdot\|C(x)\|}{n\|\delta_{R\cup C}(x)\|},\ H^{x}(C)=-\log\frac{\|C(x)\|}{n}.$$

Proof. $\delta_{R\cup C}(x)=\delta_R(x)\cap C(x)$, and we have that $\delta_R(x)\subseteq C(x)$ if x is consistent. In this case

$\delta_{R\cup C}(x)=\delta_R(x)$ .Then $-\log\frac{\|\delta_R(x)\|\cdot\|C(x)\|}{n\|\delta_{R\cup C}(x)\|}=-\log\frac{\|\delta_R(x)\|\cdot\|C(x)\|}{n\|\delta_R(x)\|}=-\log\frac{\|C(x)\|}{n}$.

Theorem 6 U is a samples set described by the features set F . $R\subseteq F$ is a subset of features and C is the decision attribute. $NMI_{\delta}(R;C)=H(C)$ if the decision of samples in feature subspace R are δ -neighborhood consistent.

Proof. As the decisions of samples in feature subspace are consistent, the decision of each sample is consistent. For $\forall x_i\in U$, $NMI_{\delta}^{x_i}(R;C)=H^{x_i}(C)$ .We have $\sum_{i=1}^{n}NMI_{\delta}^{x_i}(R;C)=\sum_{i=1}^{n}H^{x_i}(C)$.

$\sum_{i=1}^{n}NMI_{\delta}^{x_i}(R;C)=NMI_{\delta}(R;C)$, $\sum_{i=1}^{n}H^{x_i}(C)=H(C)$ .So the conclusion is correct.

Theorem 6 shows that the mutual information between features set R and decision C is equal to the uncertainty quantity of decision if the classification is consistent with respect to the attributes of R .There is not any uncertainty in classification if R is known.

3.3 Feature Selection Based On Neighborhood Mutual Information and Forward Greedy Search Strategy

Generally speaking, according to the concept of neighborhood mutual information, we can calculate neighborhood mutual information of each conditional attribute and decision attribute, and then top-ranked features are selected according to neighborhood mutual information. This method can obtain some features with higher neighborhood mutual information, but these features may contain some redundant features, and it decreases algorithm efficiency and classification performance. So a feature selection algorithm based on neighborhood mutual information and forward greedy search strategy is constructed as follows.

Algorithm 1: Feature selection based on neighborhood mutual information and forward greedy search strategy (NMI_FGS)

```
Input: samples set $U=\{x_1,x_2,...,x_n\}$ ,features set $F=\{f_1,f_2,\dots f_m\}$ ,decision attribute $C$ ,
radius of
  neighborhood $\delta$ and the threshold of termination condition $\xi$ .
  Output: a reduct $red$ .
  Step 1: $red=\phi$;
  Step 2: For each $f_i \in F-red$
      (1) calculate $NMI_\delta(f_i \cup red;C)$;
      (2) calculate $Err(f_i,red,C)=NMI_\delta(f_i \cup red;C)-NMI_\delta(red;C)$;
       End
  Step 3: Choose feature $f_k$ which satisfies: $Err(f_k,red,C)=\max_i(Err(f_i,red,C))$;
  Step 4: If $Err(f_k,red,C)>\xi$
  (1) $red=red \cup f_k$;
  (2) goto Step 2;
      else
  return and output $red$ ;
      End
```

4. Differential Evolution Algorithm

Differential evolution algorithm (DE), proposed by Stom R in1995, is a parallel and random search algorithm based on differences among groups [25, 26]. Compared with the traditional evolutionary algorithm, DE adopts real encoding to decrease complexity of genetic operation, especially its memory function has a ability of the dynamic tracking the current search condition.DE has advantages of simple principle, less parameters, better global convergence. Moreover many studies show that DE has faster convergence speed than genetic algorithm, particle swarm algorithm and ant colony algorithm, and it is easy to obtain the global optimal solution and widely is used in the optimization field.

According to the idea of neighborhood mutual information, each point in real space generates a d neighborhood, and the d neighborhood becomes a basic information particle to describe arbitrary concept of the space. The size of classification boundary is not only connected with feature space of classification problem, but also with analysis granularity, which is the size of neighborhood, namely radius. The size of neighborhood reflects the size of classification granularity and determines the amount of training samples in classification boundary region. So, the radius of neighborhood is an important factor to affect performance of neighborhood mutual information.

So far, there is no uniform standard for selecting radius of neighborhood, moreover its size is often associated with the research object, therefore radius usually is obtained by experiment method, but this method has low efficiency, and is often unable to obtain the optimal radius. In order to assure and improve the whole performance of the algorithm, this paper uses differential evolution algorithm to optimize radius of neighborhood, and steps of the optimization algorithm is as follows.

Algorithm 2:Optimization of radius of neighborhood mutual information based on DE

```
    Input: training set $X_{train}$ , testing set $X_{test}$ , population size NP, upper boundary and lower
  boundary of individual $x_{upper}$ and $x_{low}$ , mutation operator F, crossover operator CR and the
  largest number of iterations T.
```

Output: optimal radius d.

Step 1: The initialization of population.

Randomly generate a initial population $POP_0 = \{x_{i,0} \mid x_{i,0} = x_{low} + rand \times (x_{upper} - x_{low}), (i = 1, 2, \ldots, NP)\}$, where $x_{i,0}$ is a real vector whose length is 1. Each vector represents an individual in the population, Namely the radius d of neighborhood.

Step 2: Calculate individual's fitness in current population.

Training set X_{train} is reduced by using NMI_FGS with $x_{i,0}$ as radius, and then SVM is trained in training set reduced, and classification accuracy is known as the individual fitness, which is $f_{i,0} = f(x_{i,0})$.

Step 3: *For G =0 to T-1*

(1) Mutation operation: individuals $v_{i,G+1}$ are generated according to $v_{i,G+1} = x_{i1,G} + F \times (x_{i2,G} - x_{i3,G})$;

(2) Crossover operation: individuals $u_{i,G+1}$ are generated according to

$$u_{i,G+1} = \begin{cases} v_{i,G+1}, if\ rand \le CR \\ x_{i,G+1}, if\ rand > CR \end{cases};$$

(3) Calculate individual fitness $f_{i,G+1}$ in temporary population, which is $f_{i,G+1} = f(u_{i,G+1})$;

(4) Individuals $x_{i,G+1}$ are selected based on greedy criterion from $u_{i,G+1}$,and then new generation population $POP_{G+1} = \{x_{i,G+1}, i = 1, 2, ...NP\}$ is generated ;

End

Step 4: Output the optimal individual, namely radius d.

5. Our Proposed Method

Microarray data also contains many irrelevant and redundant genes which also reduce classification performance, and this paper designs a hybrid method for selecting feature genes. This method includes two stages: (1) we first use ReliefF algorithms to select a subset of top-ranked genes that have high relevance with classification task; (2) feature genes are selected from the above gene subset by using NMI_FGS. In this stage, because radius of neighborhood greatly affects the performance of NMI_FGS, so we first use differential evolution algorithm to optimize radius of neighborhood.

Our proposed method: A novel hybrid feature gene selection method

Step1: standardize data;

Step2: preselect a candidate subset of top-ranked genes based on reliefF algorithm;

Step3: optimize radius of neighborhood by using differential evolution algorithm ;

Step4: select feature genes by using NMI_FGS with optimal radius.

Step5: classify microarray datasets to verify the effectiveness of our method by using RBF-SVM as classifier.

6. Experimental Results and Analysis

6.1 Experimental Datasets and Methods

In order to evaluate the performance of our proposed method, six well-known cancer microarray datasets are selected and implemented .The characteristics of these datasets are shown in table 1. In addition, it is well known that reliefF and kruskal-wallis are better in gene selection methods filter based, so this paper compares our proposed method with reliefF and kruskal-wallis. Especially, mutual information (MI) is computed to further illustrate effectivity of our proposed method. In order to compute the mutual information of continuous features, we use a discretizing algorithm to transform these features into discrete ones [27]. RBF-SVM is adopted as the classifier in our experiment.

Table 1. Six Microarray Datasets

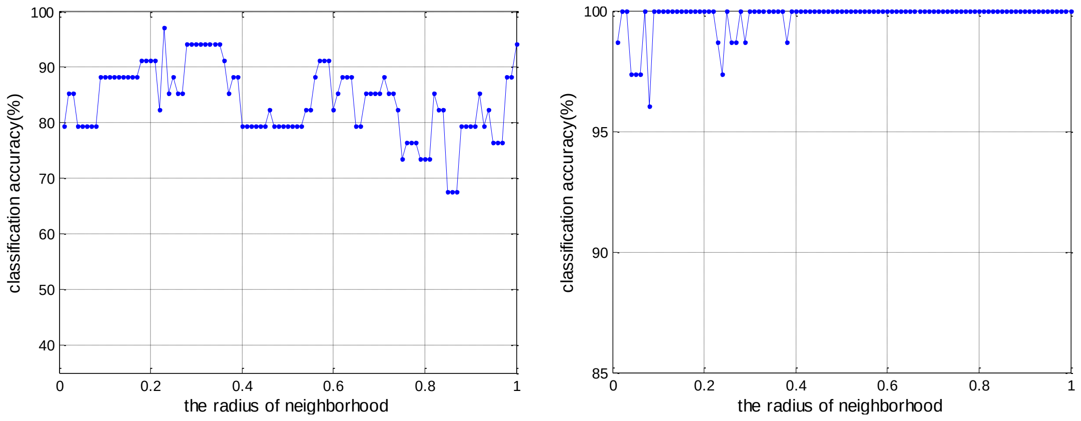
Data set	clas	ge	sam	training	testing
LeukemiaGloub(2	2	71	72	38	34
Ovarian	2	15	253	177	76
LeukemiaGloub(3	3	71	72	38	34
MLLLeukemia	3	12	72	27	45
Lung	5	12	203	142	61
DLBCL	2	71	77	32	45

6.2 Experimental Results and Analysis

6.2.1 Experiment 1: The influence of radius of neighborhood on classification accuracy based on NMI_FGS.

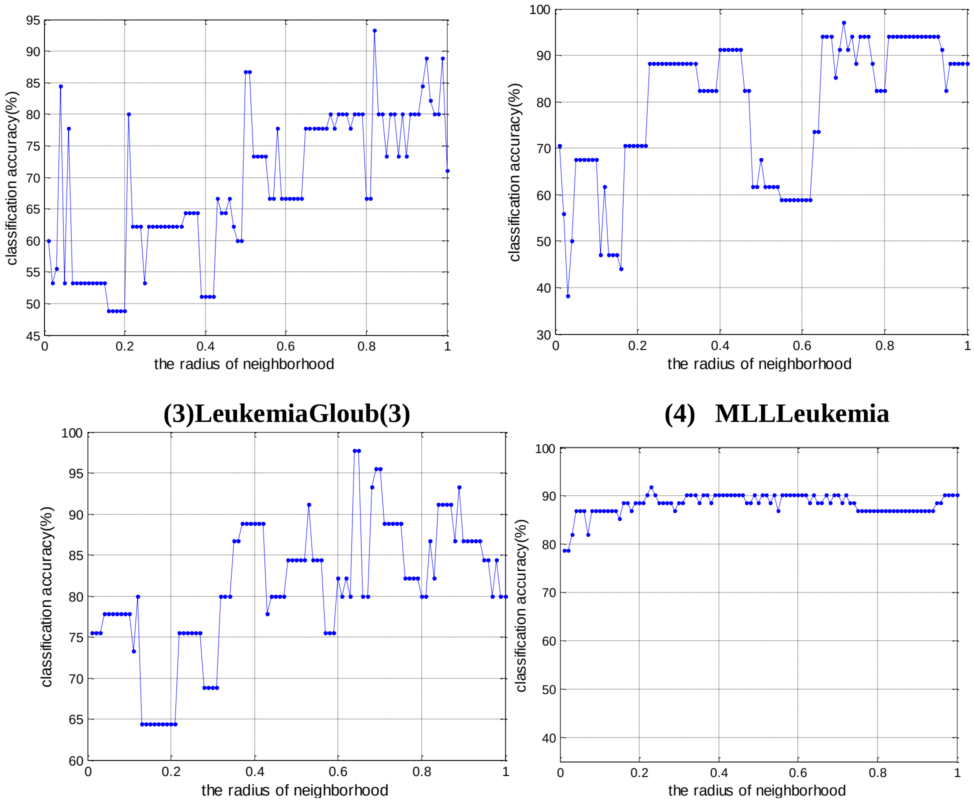
The radius of neighborhood is an important factor to affect the performance of NMI_FGS. It will obtain different classification accuracy by using NMI_FGS with different radius. In order to explain this phenomenon and analyze the influence of radius of neighborhood to classification accuracy, we will implement experiment: when taking a radius from [0,1] and the step length is 0.01, a genes subset will be got by using NMI_FGS ,and then 100 genes subsets will be generated through this method.

Figure 1 intuitively displays the influence of radius of neighborhood to classification accuracy. We clearly see classification accuracy is different by using different radius and the influence is large. Moreover if radius is inappropriate, classification accuracy will reduce greatly. In addition, we cannot accurately obtain optimal radius if we apply this experimental method, and this leads to a limitation for widespread using neighborhood mutual information. In order to guarantee performance of neighborhood mutual information, this paper applies differential evolution algorithm to optimize radius of neighborhood



(1) LeukemiaGloub

(2) Ovarian



(3)LeukemiaGloub(3)

(4) MLLLeukemia

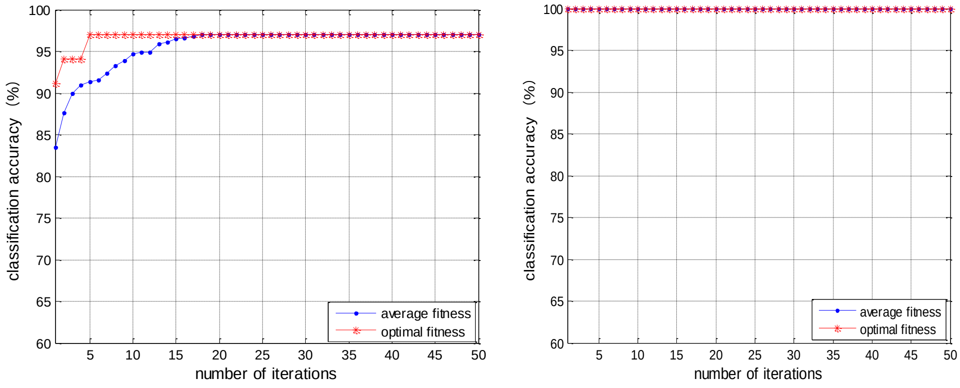
(5) Lung

(6) DLBCL

Figure 1. The Influence of Radius of Neighborhood Mutual Information to Classification Accuracy

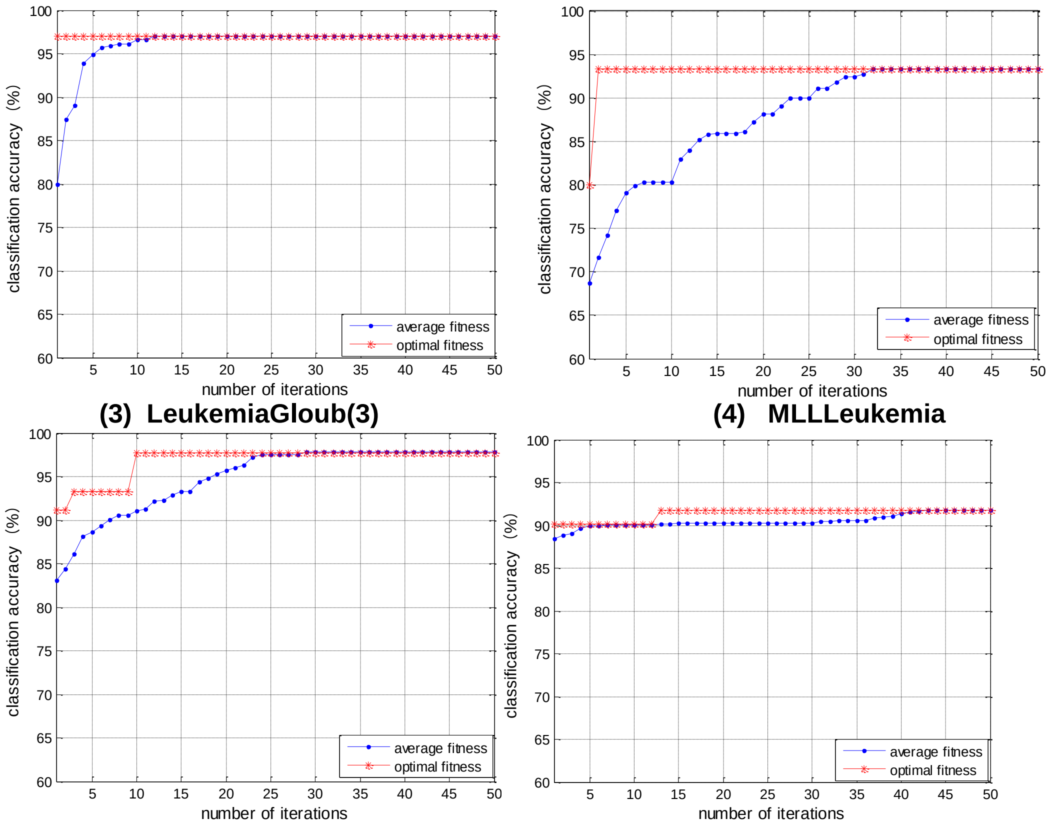
To use DE to optimize radius, some parameters of DE is set in advance according to literature [25]: population size NP=15, mutation operator F=1, crossover operator CR=0.5,the largest number of iterations T=50.

Compared with above experimental method, we can obtain optimal radius that will get the best classification accuracy by using DE algorithm. Figure 2 shows DE is an effective method for finding optimal radius, which can obtain the best classification accuracy. Moreover, the optimal radius can be obtained when iteration times is no more than 50, and it shows DE has strong global search ability and search efficiency. It is an effective and feasible method for optimizing the radius of neighborhood mutual information.



(1) LeukemiaGloub(2)

(2) Ovarian



(3) LeukemiaGloub(3)

(4) MLLLeukemia

(5) Lung

(6) DLBCL

Figure 2. The Curve of Iterative Convergence Based On DE

6.2.1 Experiment 2: Experimental results of different methods.

(1) Classification accuracy

Table 2 gives the classification accuracy of different methods on six datasets. We find that classification accuracy of our proposed method is the highest in the five methods.

Comparison with raw method, the classification accuracy of our proposed method is improved at least 38%, 19%, 41%, 24%, 16% and 21% respectively on the six datasets. On LeukemiaGloub(2), LeukemiaGloub(3), MLLLeukemia and DLBCL, result of our method is improved approximately 0.01%,3%,2% and 24% than Kruskal-wallis, respectively. For the other two datasets, results of two methods are equal. Accuracy of our method is respectively enhanced 1.5%, 6%,2%,5% and 9% than ReliefF except LeukemiaGloub(2), where results of two methods are equal. In addition, results of Kruskal-wallis and ReliefF are basically similar on six datasets.

Classification accuracy of our proposed method is obviously better than MI .It is because NMI can directly deal with continuous features to avoid information loss because of discretization. Therefore , NMI is effective method to measure relevance between continuous features and discrete decision attribute, and it can obtain higher classification accuracy than mutual information.

These results indicate our method can remove many irrelevant and redundant genes to improve classification performance. Therefore, it is very effective method for feature gene selection, especially dealing with continuous features

Table 2. Classification Accuracy of Different Methods

Dataset	Raw	Kruskal-	Relie	MI	Our	Rad
Leukemia	58.82	97.05%	**97.06**	**97.06**	**97.06%**	0.23
Ovarian	80.23	**100%**	98.68	98.68%	**100%**	0.29
Leukemia	55.88	94.11%	91.17	92.25%	**97.05%**	0.70
MLLLeuk	68.89	91.11%	91.11	91.96%	**93.33%**	0.85
Lung	75.41	**91.80%**	86.88	**91.80**	**91.80%**	0.50
DLBCL	75.56	73.33%	88.89	90.02%	**97.78%**	0.65

(2) The number of feature genes selected

Table 3 gives the number of feature genes selected of different methods on six datasets. We clearly see that the number of feature genes selected by our proposed method is far less than the number of original genes, kruskal-wallis, relief and MI. These results show our proposed method can remove irrelevant and redundant genes as far as possible and gain smaller genes subset. It decreases space dimensionality and complexity of the microarray data. The smaller number of genes can greatly improve the classification efficiency, and can enhance the understanding of microarray data. Feature gene selected of our proposed method will provide more reliable basis and reference for clinical diagnosis and pathogenic mechanism of disease.

Table 4 gives feature genes selected by using our proposed method, and the feature genes selected can help to understand microarray genes data, especially providing support for researching cancer from genomics.

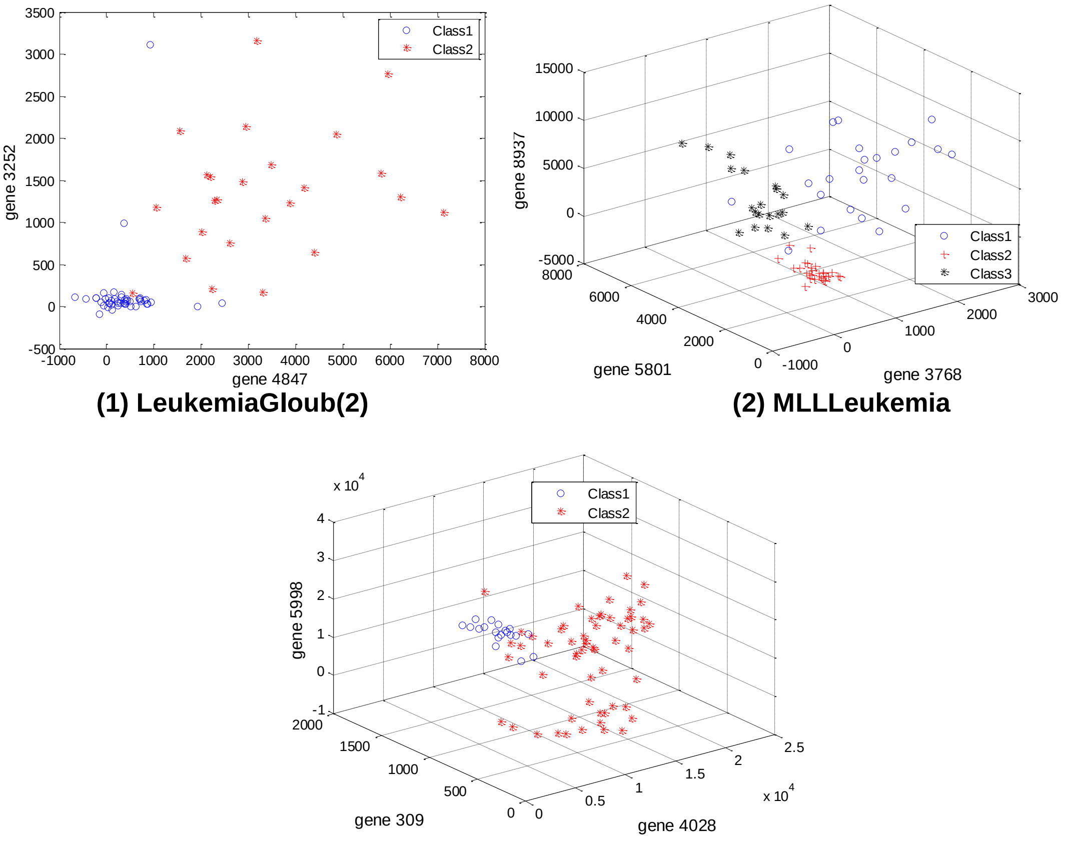
Table 3. The Number of Feature Genes Selected by Using Our Proposed Method

Dataset	Raw	Kruskal-	Relief	MI	Our method
LeukemiaGloub(2)	7129	60	20	15	**2**
Ovarian	1515	10	10	10	**7**
LeukemiaGloub(3)	7129	10	30	12	**4**
MLLLeukemia	1258	160	130	150	**3**
Lung	1260	30	40	40	**12**
DLBCL	7129	20	20	20	**3**

Table 4. Feature Genes Selected by Using Our Proposed Method

DataSet	Feature gene ID
LeukemiaGloub(2)	{4847,3252}
Ovarian	{1680,542,2237,1737,2241,2315,1683}
LeukemiaGloub(3)	{4050,2642,5876,3320}
MLLLeukemia	{3768,5801,8937}
Lung	{10139,2478,12097,4452,4786,3919,8130,10573,3600,3278,6615,12511}
DLBCL	{4028,309,5998}

Figure 3 intuitive displays scatter plot of samples in feature genes subspace on LeukemiaGloub(2), MLLLeukemia and DLBCL. From Figure 3,we clearly see that the boundary of different categories samples is clear, and it indicates only small number of feature genes can basically correctly classify complex samples, such as types of cancer. It shows that feature genes selected is effective for classifying microarray data.



(1) LeukemiaGloub(2)

(2) MLLLeukemia

(3) DLBCL

Figure 3. Scatter Plot of Samples in Feature Genes Subspace

According to above analysis, performance of our proposed method is obviously superior to reliefF, kruskal-wallis and MI. It can improve classification accuracy while the number of feature genes is decreased. It shows that the method can eliminate irrelevant and redundant genes to reveal the essential characters and structure of the high dimensional microarray data, and further improve the classification accuracy. Therefore the method is effective and efficient for selecting feature gene.

7. Conclusion

Microarray data contains a lot of irrelevant and redundant genes, feature gene selection plays an important role in improving classification accuracy and understanding microarray data .This paper presents a novel gene selection method based on ReliefF and neighborhood mutual information. ReliefF can reduce irrelevant genes from original data and feature genes are obtained by using neighborhood mutual information. Our proposed method can directly deal with continuous data, not requiring discretization. Experiment results show our method obtains higher classification accuracy and only few genes are selected. It indicates it is effective for selecting feature gene in microarray data, and selected feature genes is an important meaning for clinical research and pathogenic mechanism of cancer.

Acknowledgements

This paper is supported by National Natural Science Foundation of China (NO: 11305097), Scientific Research Program Funded by Shaanxi Provincial Education Department (NO: 14JK1148) and Scientific Research Program Funded by Shaanxi University of Technology (NO: SLGKY14-03).

References

[1] M.B.Kursa, “Robustness of random forest-based gene selection methods,”BMC bioinformatics, vol.15, no.1, **(2014)**, pp.1-8.

[2] T. Chen, “Classification algorithm on gene expression profiles of tumor using neighborhood rough set and support vector machine”,Advanced Materials Research. vol.850-851,**(2013)**,pp.1238-1242,

[3] T. Chen, “A selective ensemble classification method on microarray data”, Journal of Chemical and Pharmaceutical Research,vol.6,no.6,**(2014)**,pp.125-139

[4] Y.H Wang, F.S.Makedon, J.C.Fordand and J.Pearlman, “HykGene:a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data,” Bioinformatics,vol.21,no.8,**(2005)**,pp.1530–1537.

[5] S.L.Wang, Y.H.Zhu, W.Jia and D.S.Huang, “Robust classification method of tumor subtype by using correlation filters,” IEEE/ACM transactions on computational biology and bioinformatics, vol.9, no.2, **(2012)**, pp. 580-591.

[6] S.Zhu, D.Wang and K.Yu, “Feature selection for gene expression using model-based entropy,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol.7, no.1,**(2010)**,pp.25-36.

[7] Y. Su, T.Murali, V.Pavlovic, M.Schaffer and S.Kasif, “RankGene: identification of diagnostic genes based on expression data,” Bioinformatics, vol.19, no12, **(2003)**, pp.1578–1579.

[8] T. Li, C. Zhang and M. Ogihara, “A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression,” Bioinformatics, vol.20, no.15, **(2004)**, pp.2429–2437.

[9] 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,vol.13,**(2002)**,pp.51–60

A. BenDor,L.Bruhn,N.Friedman, I. Nachman, M.Schummer and Z.Yakhini, “Tissue classification with gene expression profiles,” Journal of Computational biology,vol.7,no.3-4,**(2000)**,pp.559–583.

[10] E.P. Xing, M.I. Jordan and R.M.Karp,“ Feature selection for high-dimensional genomic microarray data,” In Proceedings of the Eighteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc, San Francisco, USA, **(2001)**,pp.601–608.

[11] K.Kira and L.A.Rendell.“A practical approach to feature selection,” Proceedings of the ninth international workshop on Machine learning,” Morgan Kaufmann Publishers Inc, USA,**(1992)**, pp. 249-256.

[12] Kononenko, “Estimating attributes: analysis and extensions of RELIEF,” Proceedings of the European conference on machine learning, Lecture notes in computer science, **(1994)**, pp.784:171-182.

[13] H.Peng,F.Long and C.Ding, “Feature selection based on mutual information: criteria of max-dependency, max-relevance, and minredundancy,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol.27,no.8, **(2005)**,pp.1226-1238.

[14] X.Liu, A.Krishnan and A.Mondry, “An entropy-based gene selection method for cancer classification using microarray data,” BMC bioinformatics, vol.6, no.1, **(2005),** pp.126-138.

[15] R.Battiti, “Using mutual information for selecting features in supervised neural net learning,” Neural Networks, IEEE Transactions on, vol.5, no.4, **(1994)**, pp.537–550.

[16] X. Liu, A. Krishnan, A. Mondry, “An entropy based gene selection method for cancer classification using microarray data,” BMC Bioinformatics, vol. 76, no. 6, **(2005)**, pp.256–286.

[17] M.A. Hall, “Correlation-based feature subset selection for machine learning,” PhD thesis, Department of Computer Science, University of Waikato, Hamil- ton, New Zealand**,(1999)**.

[18] N. Kwak and C.H.Choi, “Input feature selection by mutual information based on Parzen window,” IEEETransactions on Pattern Analysis and Machine Intelligence, vol.24, no.12, **(2002)**, pp.1667-1671.

[19] Z. Liu and S.Wang, “From parzen window estimation to feature extraction: a new perspective,” CAAI Transactions on Intelligent Systems, vol.6, no.2, **(2012)**, pp.241-259.

[20] L.Yu and H.Liu, “Efficient feature selection via analysis of relevance and redundancy,”Journal of Machine Learning Research, vol.5, **(2004)**, pp.1205-1224.

[21] Q. Shen and A.Chouchoulas, “A rough-fuzzy approach for generating classification rules,” Pattern Recognition, vol.35, no.11, **(2002)**, pp.2425-2438.

[22] Q.H.Hu, W. Pan and S.An, “An efficient gene selection technique for cancer recognition based on neighborhood mutual information,” International Journal of Machine Learning and Cybernetics, vol.1, no.2, **(2010)**, pp.63-74.

[23] Q.H.Hu, L.Zhang and D.Zhang, "Measuring relevance between discrete and continuous features based on neighborhood mutual information," Expert Systems with Applications, vol.38, no.9, **(2011)**, pp. 10737-10750.

[24] S.Das and P.N.Suganthan, "Differential evolution: a survey of the state-of-the-art,"Evolutionary Computation, IEEE Transactions on, vol.15, no.1, **(2011)**, pp.4-31.

[25] F.I. De,"Differential evolution for automatic rule extraction from medical databases, "Applied Soft Computing,vol.13,no.2, **(2013)**,pp.1265-1283.

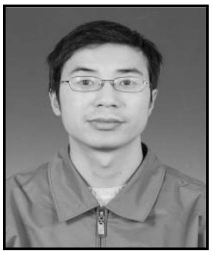
[26] U.Fayyad and K.Irani, "Multiinterval discretization of continuous valued attributes for classification learning," In Proceedings of 13th international joint conference on artificial intelligence.**(1993)**,pp.1022-1027.

Authors

Tao Chen received his Master in the school of mathematics and statistics from Lanzhou University (2007), and received his B.S. degree from Shaanxi University of Technology (2001), respectively. He is currently pursuing the Ph.D.degree in the School of Automation at Northwestern Polyechnical University. He is currently an associate professor in the school of mathematics and computer science, Shaanxi University of Technology. His current research interests are focused on data mining, pattern recognition and computational biology.

Zenglin Hong received his Ph.D.degree from School of Automation at Northwestern Polyechnical University. He is currently professor with School of Automation at Northwestern Polyechnical University. His current research interests are focused on systems engineering, land resources management and the regional economy.

Hui Zhao received his M.S. in Computer sciences (2011) from Northwest University. Now he is full instructor of informatics at school of mathematics and computer sciences department, Shaanxi University of Technology. His current research interests include different aspects of network security and data mining.



Xiao Yang received his B.S. degree from Hunan University (2002). He is currently an engineer in the school of mathematics and computer science, Shaanxi University of Technology. His current research interests are focused on data mining, pattern recognition and computational biology.

Jun Wei received his B.S. degree from Lanzhou University (2002). He is currently an experimenter in the school of mathematics and computer science, Shaanxi University of Technology. His current research interests are focused on data mining, pattern recognition and computational biology.