A Feature Gene Selection Method Based on ReliefF and PSO

Liu Mengdi [1,2], Xu Liancheng[1,2,*], Yi Jing[1,3], Huang Jie[1,2]

(1.School of Information Science and Engineering, Shandong Normal University, Jinan,Shandon-g,250358,China; 2.Shandong Provincial Key Laboratory for Distributed Computer Software No-vel Technology,Jinan,Shandong,250358,China;3.School of Computer Science and Technology, Shandong Jianzhu University, Jinan, Shandong, 250014, China)

*corresponding author's email: lemonly0209@163.com

***Abstract*—Due to the shortcomings of low efficiency and low accuracy of DNA microarray data, the paper proposes a feature gene selection algorithm based on ReliefF and PSO (RefFPSO). Firstly, ReliefF is used as the feature pre-filter to delete the genes with low correlation with the classification target. Then PSO is used as the search algorithm. Finally, the classification accuracy of SVM is used as the evaluation function of the feature subset to get the final optimal gene subset . Experiments show that this method can effectively put forward irrelevant genes and use fewer characteristic genes to obtain higher classification accuracy.**

Keywords- DNA microarray data; ReliefF algorithm; particle swarm optimization; feature selection

I. INTRODUCTION

Microarray technology is a rapidly developing molecular biology technology and is one of the major technologies of gene revolution. The most prominent advantage of DNA microarray technology is that it can detect a variety of samples at once and realize rapid detection of genetic information. Microarray technology has gained widespread use in environmental monitoring, drug discovery and evolutionary biology. There are many ways to classify microarray data. The most commonly used method is to use machine learning and data mining techniques to build a model to classify samples and extract valuable information. The DNA microarray data exist high-dimensional data, the characteristics of small samples, it is easy to "dimension explosion" problem, which had a negative impact on the classification results. However, in the actual research, only a few gene features are involved in the classification decision when classifying large data of DNA microarray. Most of the genes are unrelated to the classification target or have little correlation. Therefore, the data pretreatment, screening out the necessary genes, gene expression data classification is to achieve accurate and efficient necessary steps.

The method of feature gene selection is to find the most effective subset of genes from the original genome so as to reduce the classification cost and improve the classification accuracy. The method of feature selection can be divided into Filter and Wrapper in accordance with the working principle of evaluation function [1,2]. The Filter method is independent of a specific model and has a fast selection speed but a large scale of its own. The Wrapper method needs to be combined with a specific model. The selected feature set usually has a high classification accuracy, but it is difficult to process the high-dimensional feature set and calculate Great cost. How to combine Filter and Wrapper to further improve the classification performance is the key point of feature selection algorithm. The former is the feature pre-filter, and the latter is based on the pre-screening feature set for further feature selection [3]. Qiu Guoyong et al. [4] proposed a two-stage feature selection method based on standardized mutual information and genetic algorithm. The experimental results show the effectiveness of the proposed method and the classification accuracy is reduced while reducing the number of features.

Relief is a well-known Filter feature selection method. Relief is a series of algorithms, including the earliest proposed Relief and later developed Relief and ReliefF. The earliest proposed Relief is for dichotomous problems, ReliefF algorithm can solve multi-classification, data missing And the existence of noise and other issues [5]. Relief series algorithms have low time complexity and are algorithms based on feature weights, so the redundant features can not be effectively removed by excluding features with low weight value. Classifiers and search algorithms are two components of Wrapper feature selection. SVM has become a hotspot in pattern recognition and machine learning in recent years due to its advantages of small sample size, strong anti-noise performance, high learning efficiency and good generalization [6-7] A research hotspot in the field that can be effectively used for Wrapper feature selection; Particle swarm optimization (PSO) has attracted the attention of academics for its excellent global search optimization capability, and has become one of the main search algorithms for encapsulation with SVM.

In summary, this paper presents a method of feature gene selection based on ReliefF and PSO (RefFPSO). Firstly, ReliefF is used as a feature preprocessor to filter out some of the less relevant gene features. Then using PSO as the search algorithm, the SVM classification accuracy As the evaluation function, the best feature gene subset is selected among the remaining gene features. Finally, experiments were carried out on several common gene expression datasets. Experiments show that this method can effectively remove irrelevant and redundant genes from taxonomic data and improve the classification accuracy.

II. RELIEFF ALGORITHM

Relief algorithm as a high-efficiency Filter algorithm, which according to the importance of the order of the features, and will be higher than the specified threshold as a feature subset. The theoretical basis is that a good feature should make the eigenvalues of similar samples in the nearest neighbor be close to each other, and make the

2157-1481/18/$31.00 ©2018 IEEE
DOI 10.1109/ICMTMA.2018.00079

difference between the different samples in the nearest neighbor be large. According to this, the feature weights are given to each feature for feature ordering. The greater the feature weight, the stronger the classification ability of the feature. On the contrary, it indicates that the feature classification ability weaker. The corresponding feature selection can be done by setting the thresholds of feature weights or the number of feature subsets. Relief by (1) update the weight of the feature A,W [A]:

$$W[A] = W[A] - diff(A,R,H)/m + diff(A,R,H)/m \quad (1)$$

Where R is the sample randomly selected in the training set, H and M are respectively the nearest and neighbor samples of the same type and the same type of sample R, m is the number of random samples, and the diff function represents the difference between the two samples of the given attribute . When calculating W [A], normalize it with m to ensure that the weight value is between -1 and 1.

When the feature attribute is a nominal property, *diff (A, I_x, I_y)* is calculated as follows [8]:

$$diff(A,I_x,I_y) = \begin{cases} 0; value(A,I_x) = value(A,I_y) \\ 1; otherwise \end{cases} \quad (2)$$

When the characteristic attribute is a numerical attribute, *diff (A, I_x, I_y)* is calculated as follows:

$$diff(A,I_x,I_y) = \frac{\left| value(A,I_x) - value(A,I_y) \right|}{\max(A) - \min(A)} \quad (3)$$

The traditional Relief algorithm is limited to dichotomous problems and can not deal with the problem of missing data and noise. The emergence of Relief F solves the above problem by reducing the effect of noise in the data by averaging the k nearest neighbor differences of the same kind for each sample and the k different nearest neighbor differences of the other classes. The choice of nearest neighbor samples is the most fundamental difference between Relief and Relief F, which ensures the robustness of the Relief F algorithm.

III. PSO ALGORITHM

Particle swarm optimization (PSO) is an evolutionary computation technique. From the behavior of bird foraging behavior. Particle swarm optimization algorithm's basic idea is to find the optimal solution through the collaboration and information sharing among individuals in the group. PSO is initialized to a group of random particles. In a d-dimensional target search space, the spatial position of particle i in particle swarm is expressed as $X_i = (x_{i1}, x_{i2}, ..., x_{id})$ and flies i at velocity $V_i = (v_{i1}, v_{i2}, ..., v_{id})$ in search space. Particle spatial location is a solution to the problem of objective optimization. The fitness value is used to calculate the fitness value and the fitness value is used to measure the particle's fitness. Particle velocity adjusts its current spatial position, and the PSO finds the optimal solution iteratively [9]. The velocity and position of particle i in the j-th space are updated as follows:

$$v_{ij}(t+1) = w \cdot v_{ij}(t) + c_1 \cdot rand() \cdot (p_{ij}(t) - x_{ij}(t)) + c_2 \cdot rand() \cdot (p_{gj}(t) - x_{ij}(t)) \quad (4)$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \quad (5)$$

Where t represents the number of iterations, w is the inertia coefficient, rand () is a random number between 0 and 1, and $p_i = (p_{i1}, p_{i2}, ..., p_{id})$ is the local optimum position that the particle i passes in the search space , $Pg = (p_{g1}, p_{g2}, ..., p_{gd})$ is the global optimum of the whole particle swarm, and c1 and c2 are constant coefficients, which are used to adjust the local search ability. According to the literature [10], c1 and c2 both take values of 1.42694 .

In order to meet the needs of discrete problems, Kennedy and Eberhart [11] proposed a binary-coded particle swarm optimization algorithm, which binary coded every particle, and the particle velocity represented the particle position assignment as 1 probability, it is necessary to convert the speed function to the [0,1]. This article uses the sigmoid function for conversion, particle location update formula is as follows:

$$x_{ij} = \begin{cases} 1, rand() \le sig(v_{ij}) \\ 0, otherwise \end{cases} \quad (6)$$

$$sig(v_{id}^{n+1}) = \frac{1}{(1 + \exp(-v_{id}^{n+1}))} \quad (7)$$

Where $s(v_{ij})$ denotes the probability that the value of x_{ij} is 1.

The fitness function is an important index for the PSO algorithm to evaluate the advantages and disadvantages of the feature subsets, usually using the classification learning algorithm. However, in high-dimensional data sets, the use of classification learning algorithm as a fitness function often leads to higher time and memory overhead. Inconsistency rate as a measure of consistency, you can measure the pros and cons of feature sets. Inconsistent rate with monotonous, low computational complexity, the ability to remove redundant and irrelevant features. Therefore, this paper chooses the inconsistency as the adaptive function of PSO algorithm.

$$fitness = \frac{N_{in}}{P} \quad (8)$$

Where P is the total number of samples; we call a combination of features in a sample instance a pattern, and the number of inconsistencies Nin for all patterns in the feature subset equals to the total number of samples appearing in the pattern minus the number of occurrences The number of labels.

IV. A FEATURE SELECTION METHOD BASED ON RELIEFF AND PSO

According to the characteristics of Relief F algorithm and PSO algorithm, this paper proposes a feature gene selection algorithm based on ReliefF and PSO (RefFPSO). The algorithm consists of two parts: Firstly, the weight of each feature gene is calculated by ReliefF algorithm, and then the features are sorted, and the feature combinations

with stronger discriminative power and higher discriminative power are selected as the candidate gene subsets; then the candidate genes As a subset of the input of the PSO algorithm, the subset is gradually removed redundant features in the iterative process to get the optimal gene subset.

The flow chart of the algorithm is as follows:

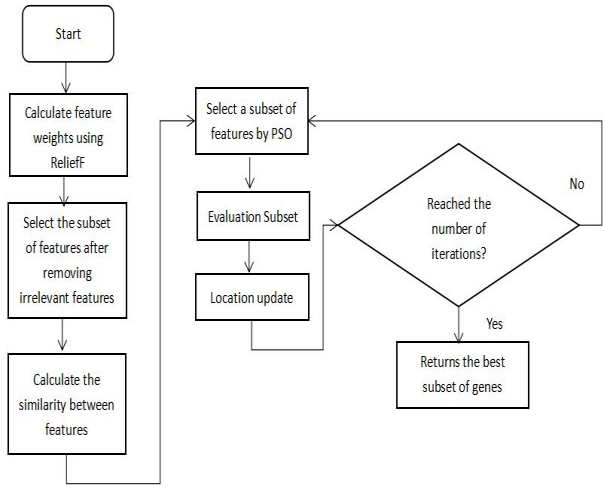

Figure 1 Algorithm flowchart

V. EXPERIMENTAL RESULTS AND ANALYSIS

In order to evaluate the performance of this algorithm in DNA microarray data processing, it is used in four gene data sets of colon cancer, SRBCT, leukemia and lung cancer. The data set is briefly described in Table 1. The algorithm used in this experiment is programmed by Matlab7. The CPU of the computer used in this experiment is Intel Core i3, clocked at 3.30GHz, memory is 2GB and operating system is Windows 7. The classic classifier used in this paper is SVM.

Table 1 Gene data sets

DataSet	Gene	Sample	Class
Colon cancer	2000	62	Normal(22),Tumor(40)
SRBCT	2308	83	EWS(29),BL(11), NB(18),RMS(25)
Leukemia	7129	72	ALL-T(9),ALL-B(38), AML(25)
Lung cancer	12600	203	Adeno(139),NORM(17) Squamous(21) COID(20),SMCL(6)

The experiment is mainly divided into two parts: the first part compares the number of features selected by different feature selection methods; the second part compares the effect of the selected feature genes on the classification results under different methods.

A.Comparison of feature selection methods

In order to fully prove the effectiveness of the proposed algorithm, this group of experiments choose four commonly used feature gene selection method with the proposed method for comparison. Among them,FCBF [12] CFS [13],mRMR-ReliefF[14].

Table 2 The number of features under different algorithms

	ReliefF	mRMR ReliefF	CFS	FCBF	RefFPSO
Colon cancer	17	23	22	9	13
SRBCT	21	26	26	17	26
Leukemia	28	31	34	12	24
Lung cancer	32	45	32	21	33
Average value	25	31	29	15	24

Table 2 shows the number of features selected by the five feature selection methods on the four data sets. It can be seen that the number of features selected by FCBF algorithm is the least, followed by the algorithm in this paper, and mRMR-ReliefF selects the largest number of features.

B.Comparison of classification accuracy

In this paper, SVM classifier to verify the classification accuracy of the above five algorithms. Figure 2 shows the accuracy of classification of the four data sets by five feature selection algorithms.

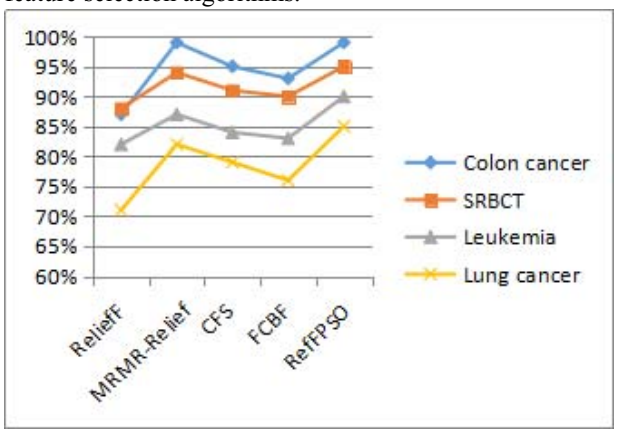

Figure 2. Classification accuracy

Figure 2 shows the classification accuracy of the five algorithms based on SVM. From the experimental results, it can be found that the classification accuracy of the features selected by the RefFPSO algorithm presented in this paper is the highest on the four data sets used, and mRMR-Relief F is the second highest. When using the SVM algorithm to classify colon cancer data, both m RMR-Relief F and Re FACO can be completely correct. The accuracy of the RefFPSO algorithm is obviously higher than the other four when classifying the other three data. And with Table 1, it is found that RefFPSO algorithm can not only solve the

multi-classification problem, but also has good results in the two classification. The RefFPSO algorithm achieves the best classification accuracy based on the relatively small number of selected features.

VI. CONCLUSIONS

In order to select effective feature combination in DNA microarray data of high dimensional small samples, this paper proposed a feature selection algorithm based on Relief F and PSO based on ReliefF and PSO. The algorithm first uses ReliefF to quickly reduce the dimension, and then initializes the particle community with the subset with larger feature weight, and uses the inconsistency as the fitness function to remove the redundant features. Experiments show that the proposed algorithm achieves better classification results with fewer characteristic genes on different scales of data sets.

ACKNOWLEDGEMENTS

This research was financially supported b-y the Natio nal Natural Science Foundation Project (No.61373148, No.61502151),Ministryof Education Humanities and So cial ScienceFoundation Project(No.14YJC860042),the Na t-ural Science Foundation of Shandong Provin-ce(No.ZR 2014FL010),Shandong Province Outstanding Young Scie ntists Award Fund fundedProjects(No.BS2013Dx033),Sh andong ProvinceHigher Education Science and Technolo gy P-rogram (No.J15LN02, No.J15LN22),ShandongProvi nce Social Science Planning Project(No.16CxWJ01, No. 16CFxJ05).

REFERENCES

[1] Yao Xu,Wang Xiaodan,Zhang Yuxi,Quan Wen,Overview of feature selection methods [J].Control and Decision,2012,27(02):161-166+192.

[2] Lorena L H, Carvalho A, Lorena A C.Filter feature selection for one-class classification [J].Journal of Intelligent & Robotic Systems, 2015,80(1):227-243.

[3] Cadenas J M,Garrido M C,Martinez R.Feature subset selection Filter-Wrapper based on low quality data[J].Expert Systems with Applications, 2013, 40(16): 6241－6252.

[4] Qiu Guoyong, Wang Na, Wang Wanzi. Two-stage feature selection algorithm based on mutual information and genetic algorithm[J]. Application Research of Computers, 2012, 29(8): 2903－2905.

[5] Reyes O, Morell C, Ventura S. Scalable extensions of the Relief F algorithm for weighting and selecting features on the multi-label learning context [J]. Neurocomputing, 2015, 161 (C): 168-182.

[6] Lin Nan,Jiang Qigang, Yang Jiajia, et al. Classifications of agricultural land use based on high-spatial ZY1-02C remote sensing images[J]. Transactions of the Chinese Society for Agricultural Machinery, 2015, 46(1): 278－284.

[7] Qi Zhixin, Yeh A G, Li Xia, et al. A novel algorithm for land use and land cover classification using RADARSAT-2 polarimetric SAR data[J].Remote Sensing of Environment, 2012, 118: 21－39.

[8] Kira K, Rendell L A. A practical approach to feature selection [C]// Proc of International Workshop on Machine Learning. 1992: 249-256.

[9] Tang Jun. Principle and application of PSO algorithm[J].Computer Technology and Development, 2010, 20(2): 213－216.

[10] QIN Quande, CHENG Shi, ZHAN Qingyu, et al.Particleswarm optimization with interswarm Interactive learningstrategy [J] . IEEE Transactions on Cybernetics, 2016, 46(10) : 2238－2251.

[11] Kennedy J, Eberhar R. A discrete binary version of the particle swarm algorithm[C].Proceeding of the World Multiconference on Systemics, Cybernetics and Informatics, Newjersy: Piscataway, 1997: 4104－4109.

[12] Hall M A. Correlation-based feature selection for discrete and numeric class machine learning [C]//Proc of the 17th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc, 2000: 359-366.

[13] Yu L,Liu H. Feature selection for high-dimensional data:A fast correlation-based filter solution [C]// Proc of the 20th International Conference Machine Learning. 2003: 856-863.

[14] Zhang Y, Ding C, Li T. Gene selection algorithm by combining relief F and m RMR [J]. BMC Genomics, 2008, 9 (suppl 2): 1-10.