A Hybrid Gene Selection Method for Microarray Data Based on Geodesic Distance and Binary Particle Swarm Optimization

IOP Conf. Series: Materials Science and Engineering **490** (2019) 042014 doi:10.1088/1757-899X/490/4/042014

A Hybrid Gene Selection Method for Microarray Data Based on Geodesic Distance and Binary Particle Swarm Optimization

Ying Xiong*, Fei Han

School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China

*Corresponding author e-mail: 2211608015@stmail.ujs.edu.cn

Abstract. To obtain the most predictive genes subsets without filtering out critical genes, a method for gene selection based on binary particle swarm optimization (BPSO) and geodesic distance is proposed in this paper. In this approach, to preserve the intrinsic geometry of high dimensional microarray data, geodesic distance is calculated as the measurement between genes for cluster, and by combining with clustering method, BPSO is used to perform gene selection to reduce redundancy. With experiments conducted on several public microarray data by ELM classifiers, the results confirm that it is efficient to use the proposed method for gene selection compared to the relevant gene selection method.

1. Introduction

Gene selection is a critical data preprocessing technique in the classification [1], which could decrease the computational complexity and gene redundancy, as well as increases the classification accuracy [2]. Though a large pool of methods are already available, selecting the best gene subset for accurate classification is still very challenging.

As a swarm intelligence optimization algorithm, binary particle swarm optimization (BPSO) [3], developed to make particle swarm optimization (PSO) [4-6] has the ability to optimize the combination problem in order to be used in discrete space, has been widely used for various researches and application areas, especially in feature selection [7-8]. In [9], to search the optimal gene subsets, BPSO combined with filter method was proposed. And in [10] a modified discrete PSO was used with support vector machines (SVM) to select genes, which demonstrated that the discrete PSO could be a powerful method for gene selection. In [11], BPSO and Bayesian Linear Discriminant Analysis (BLDA) were proposed to combine to select genes with lower redundancy and high classification accuracy. Despite the fact that these gene selection methods had a superior performance on selecting predictive gene subset to achieve high classification accuracy, but the true geometric structure of the genes was not taken full consideration in the selection process. Two gene selection methods on the basis of BPSO and gene-to-class sensitivity (GCS) information were proposed in [12][13] for not only achieving high classification accuracy but also obtaining predictive genes with more interpretability. In [13], the GCS information was encoded into an improved BPSO to select smallest gene subsets with better interpretability and lower redundancy. Nevertheless, some key genes may be ignored in some cases and thus result in lower accuracy for classification.

According to the above analysis, some current methods for gene selection based on PSO have the disadvantages that lack of considering the true geometric structure of the genes and even filter out some key genes. In this paper, to overcome the deficiencies as well as improve prediction accuracy, a hybrid method which based on geodesic combined with BPSO is proposed. Firstly, using a filter method to filter out redundant genes. Secondly, dividing the initial gene pool data into cluster by using the K-medoids approach based on geodesic distance instead of Euclidean distance. Finally, the BPSO combine with ELM for higher classification accuracy are used to select the optimal gene subset. Furthermore, the geodesic distance which indicating the relevance between genes is further considered into the BPSO fitness function. Experimental results demonstrate the proposed hybrid gene selection method is effective.

2. Preliminaries

2.1 Binary Particle Swarm Optimization

Each particle in BPSO has two attributes in D dimensional search space. The position and velocity of each particle can be represented as $X_i = (x_{i1}, x_{i2}, \dots, x_{iD})$ and $V_i = (v_{i1}, v_{i2}, \dots, v_{iD})$ $i = (1, 2, \dots, N)$ respectively, N is the population size. The features of global best position and personal historical best position could guide each particle to adjust its velocity and position according to the following equation:

$$V_{ij}(t+1) = \omega * V_{ij}(t) + c_1 * r_1 * \left(P_{ij}(t) - X_{ij}(t)\right) + c2 * r_2 * \left(G_j(t) - X_{ij}(t)\right) \tag{1}$$

$$X_{ij} = \begin{cases} 1 & rand() \le S(V_{ij}) \\ 0 & rand() > S(V_{ij}) \end{cases} \tag{2}$$

$$S(V_{ij}) = \frac{1}{1+\exp(V_{ij})} \tag{3}$$

where $j = (1, 2, ..., D)$; $P_{ij} = (P_{i1}, P_{i2}, ...P_{iD})$ and $G(t) = (g_1, g_2, ..., g_D)$ is the i-th particle's individual best position and the global best position in current iteration respectively; ω is the inertial weight of BPSO; t denotes the iteration number; $c1$ and $c2$ are two acceleration factors which can balance the impact of P_i and G_j; Both $r1$ and $r2$ are randomly generated in [0, 1].

2.2 Geodesic Distance

For nonlinear data, especially for flow distribution data, the Euclidean distance between genes does not fully reflect the relationship. Geodesic distance can truly reproduce the nonlinear geometry in high-dimensional microarray data[15]. The main steps for geodesic distance can be described as follows:

step 1. Calculate the Euclidean distance $Dis_E(X_i, X_j)$ of any two sample X_i and X_j;
step 2. Determine the K neighbors of the sample X_i;
step 3. Calculate the shortest path between any two samples X_i and X_j;
step 4. Calculate the Euclidean d
istance $Dis_G(X_i, X_i)$

$$Dis_G(X_i, X_i) = \begin{cases} Dis_E(X_i, X_j) & X_i \text{ is the } K \text{ neighbors of } X_j \\ \min_{i,j,k}\{Dis_G(X_i, X_j), Dis_G(X_i, X_k) + Dis_G(X_k, X_j)\} & otherwise \end{cases} \tag{4}$$

3. The proposed gene selection method (GD-Kmedoids-BPSO-ELM)

The proposed method is aimed to deal with the problem about how to select the best gene subsets based on the true geometric structure of the genes to improve the classification accuracy without filtering out key genes. Since the proposed method combines the BPSO with K-medoids based on geodesic distance and ELM, thus the proposed method is called GD-Kmedoids-BPSO-ELM in short. The following are the detailed steps summarized from the method：

Step 1: Initial a first-level gene pool by filter method. 200 genes are selected by using a filter method called Signal-to-Noise Ratio (SNR). Divide the dataset into two parts: training and testing datasets.

Step 2: Establish the candidate elite gene pool. To give full consideration of gene structure as well as decrease the computational complexity, the geodesic distance is calculated to show the true flow structure of genes. And the k-medoids is used to cluster the genes.

Step 3: Select the optimal gene subsets by BPSO both for high accuracy and high interpretability in gene structure. Initialize all particles. The position x_{ij} can be coded to 0 or 1, 1 means the j-th gene is selected and 0 means the j-th gene is not selected. its pbest is initialized by the current position of each particle，and the gbest is initialized by find the best pbest. The particles are updated by the computed fitness value. Moreover, the accuracy obtained by ELM and the distance between selected genes are both considered into the fitness. As shown in Eq. (5), by using a weighting coefficient (λ), the classification accuracy and a distance measure are balanced in the proposed fitness function. The geodesic distance between genes from different cluster (DB) and between genes from the same cluster (DW) are maximized for the purpose of less redundancy and more interpretability. DB and DW are updated according to Eq. (7) and (8) respectively:

$$fitness = \lambda * Accuracy + (1-\lambda) * distance \tag{5}$$

$$distance = \frac{1}{1+e^{-(D_B+D_W)}} \tag{6}$$

$$D_B = \frac{1}{|N_S|}\sum_{k=1}^{N_S} \max_{\{j|\, j\neq i, cluster(i)\neq cluster(j)\}} Dis_G(i,j) \tag{7}$$

$$D_W = \frac{1}{|N_d|}\sum_{k=1}^{N_d} \max_{\{j|\, j\neq i, cluster(i)=cluster(j)\}} Dis_G(i,j) \tag{8}$$

where the $Accuracy(i)$ is the classification accuracy at the selected gene subset denoted by the *i-th* particle from ELM. N_S is the number which genes arevfrom different cluster and N_d is the number which genes are from the same cluster respectively.

4. Experimental results and discussions

4.1 Datasets

The experiments are performed on three open microarray datasets involve Brain cancer, Lymphoma data and Lung data. The descriptions for the datasets are listed in Table 1.

The Lung data include 203 samples in five classes with 3,312 genes. In Brain cancer there are classic 46 samples and 14 desmoplastic samples with 7129 genes, a total of 60 Brain cancer samples divided into 2 categories. And in Lymphoma data, it contains 32 cured samples and 26 samples which are not cured with 7129 genes.

Table 1. Specification of three microarray datasets

Data	Total samples	Training samples	Testing samples	Number of classes	Number of genes
Brain cancer	60	30	30	2	7129
Lymphoma	58	29	29	2	7129
Lung	203	103	100	5	3312

In the experimental data, the size of the swarm is 30, the number of maximum iteration is set as 100, Both $c1$ and $c2$ is set as 1.49445, ω change start at 0.9 to 0.4. The cluster number is fixed as 5 on all data. The hide node number in ELM is 300.

Table 2. The gene subsets selected by the proposed method and corresponding classification accuracy on three microarray data

Data	Selected gene subsets	5-fold CV Accuracy Mean(%)±std	Test Accuracy Mean(%)±std
Brain	18,3341,1582,2942,1198,6331,4917,724,	92.25±0.748	90.78±1.103
	6429,	92.34±0.623	91.73±1.065
	4372 ,6774,1975,587,2122,5051,6700,6828		
	6429,4309,3555,1975,3035,3341,1648,161, 724	91.17±0.672	90.23±1.145
Lymphoma	4862,3589,3227,704 ,2810,4998	92.21±0.523	90.63±0.782
	4514,3589,5709,6172,2666,2810,3525	92.23±0.063	90.24±0.953
	3589,3775,5709,6565,5329,418,5818	92.14±0.234	91.51±0.723
	1268,1822,2356,445,2556,1318,1411,295,2005,1712	94.44±0.782	90.61±1.220
Lung	2479,924,2969,1974,1822,580,2279,2128,1411,2005,414	95.45±0.432	92.56±1.013
	1268,3276,2969,441,295,2904,445,2895,2128,261,1028, 2005	97.45±1.023	93.89±2.103

4.2. The classification ability analysis on genes selected by the proposed method.

The selected gene subsets by the proposed method and corresponding classification accuracy on three microarray data are in Table 2.

From the shown genes selected in Table 2, the GD-Kmedoids-BPSO-ELM method selects about five to twelve genes. And the corresponding classification accuracy is high. The best accuracy is 92.34%,92.23%, 97.45% on Brain, Lymphoma, Lung data respectively. The results show that the GD-Kmedoids-BPSO-ELM method is able to select those predictive genes.

4.3. The comparison with other relevant gene selection method .

The GD-Kmedoids-BPSO-ELM are compared with BPSO-ELM, Kmeans-BPSO-ELM and Kmeans-GCSI-MBPSO-ELM. for those algorithms in all experiments the parameters are determined by repeated test. The maximum optimization epochs for BPSO-ELM, KMeans-BPSO-ELM, KMeans-GCSI-MBPSO-ELM and our method are 100. The corresponding results of the average of 100 trials are listed in Table 3. The four programs are run in MATLAB 8.1 environment.

Table 3. The comparison with other three relevant gene selection method on three microarray data.

Method	5-fold CV Accuracy（Mean%±std）and selected gene number		
	Brain	Lymphoma	Lung
BPSO-ELM	85.45±2.33(7)	83.50±2.72(8)	94.80±0.57(11)
KMeans-BPSO-ELM	87.23±2.34(8)	85.14±2.87(6)	95.64±0.56(12)
KMeans-GCSI-MBPSO-ELM	88.63±2.16(6)	86.97±2.44(8)	97.10±0.63(11)
The proposed method	92.34±0.623 (9)	92.23±0.063 (7)	97.45±1.023(12)

From the results listed in Table 3, the proposed method in this study outperforms other ELM-based method on Brain and Lymphoma data. While the selected gene number almost is the same as other three method. In Lung data the accuracy is slightly higher than the KMeans-GCSI-MBPSO-ELM but still is superior to other two methods. The classification accuracy in every iteration of BPSO on three microarray data is shown in Fig.1. In the GD-Kmedoids-BPSO-ELM, the BPSO evolves 42,34,43

epochs on Brain, Lymphoma, and Lung data to selects the optimal subsets. Which shows the GD-Kmedoids-BPSO-ELM is capable to select critical gene subset and achieve the higher classification accuracy.

4.4 Discussion on the Parameter Selection

It is critical to determine the parameter of cluster number when clustering. The relationship between the accuracy and the cluster number k are depicted in Fig. 1. From the Fig.2, the accuracy when the cluster number k is higher than that when the cluster number k =1, therefore it illustrates that clustering is helpful to the gene selection process. The accuracy is highest when the cluster number k = 5 among the three dataset. Thus, setting the cluster number as 5 could make the accuracy highest.

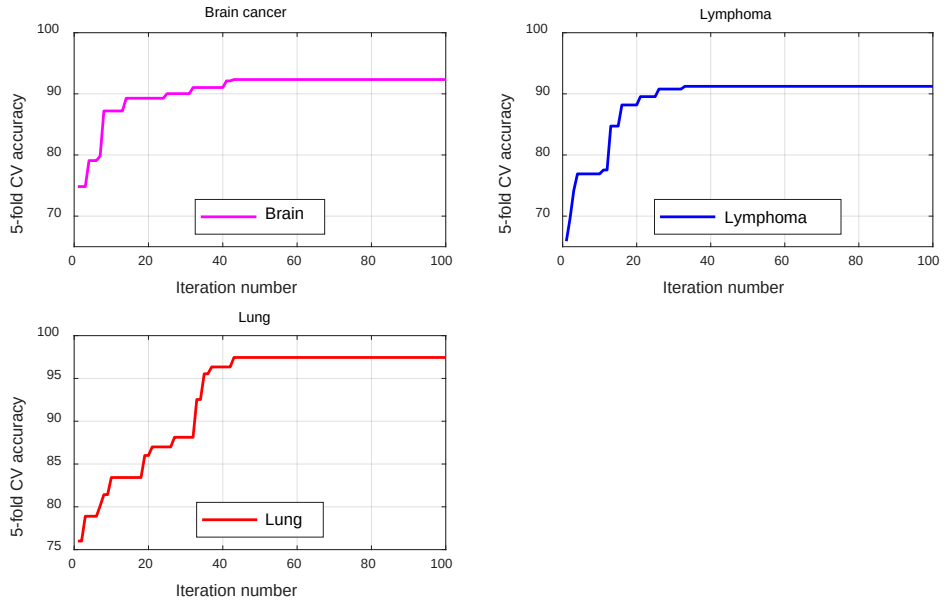

Figure 1. The iteration number versus 5-fold CV accuracy of selected genes on the three datasets. (a) Brain (b) Lymphoma (c) Lung.

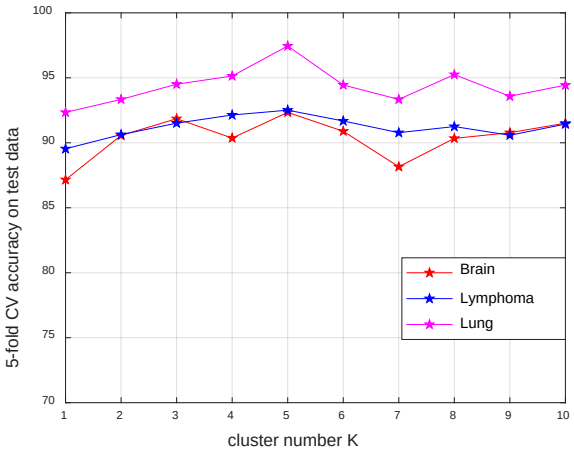

Figure 2. The cluster number versus 5-fold CV accuracy of selected genes on the three datasets

5. Conclusions

In this study, give full consideration of gene structure as well as decrease the computational complexity, initial gene pool data are divided into clusters by using the K-medoids approach based on geodesic distance. Finally the BPSO selects the optimal subsets, to obtain the most predictive genes subsets without filtering out critical genes, both classification accuracy and gene geodesic distance are

considered in the fitness function. Experimental results verified the proposed method could select highly predictive and compact gene subsets and outperformed than other PSO-based and GCSI-based gene selection methods. Future work will study applying this proposed method to more microarray data.

Acknowledgment
This work was supported by the National Natural Science Foundation of China [Nos. 61572241 and 61271385], the Foundation of the Peak of Six Talents of Jiangsu Province [No. 2015-DZXX-024], the Fifth "333 High Level Talented Person Cultivating Project" of Jiangsu Province [No. (2016) III-0845].

References

[1] Saeys Y, Inza I, Larrañaga P. WLD: review of feature selection techniques in bioinformatics. Bioinformatics, 2007, 23(19):2507-2517.
[2] Yang C S, Chuang L Y, Ke C H, et al. A Hybrid Feature Selection Method for Microarray Classification. Iaeng International Journal of Computer Science, 2008, 35(3).
[3] Kennedy J, Eberhart R C. A discrete binary version of the particle swarm algorithm, IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation. IEEE, 2002, vol.5:4104-4108.
[4] Kennedy J, Eberhart R. Particle swarm optimization, IEEE International Conference on Neural Networks, 1995. Proceedings. IEEE, 2002, vol.4:1942-1948.
[5] Mohamad M S, Omatu S, Deris S, et al. Particle swarm optimization for gene selection in classifying cancer classes. Artificial Life & Robotics, 2009, 14(1):16-19.
[6] Wang Y, Wang Y, Chen Y, et al. Particle Swarm Optimization (PSO) for the constrained portfolio optimization problem. Expert Systems with Applications, 2011, 38(8):10161-10169.
[7] Mohamad M S, Omatu S, Deris S, et al. A Modified Binary Particle Swarm Optimization for Selecting the Small Subset of Informative Genes From Gene Expression Data. IEEE Transactions on Information Technology in Biomedicine, 2011, 15(6):813-822.
[8] Liu J, Fan X. The Analysis and Improvement of Binary Particle Swarm Optimization, International Conference on Computational Intelligence and Security. IEEE Computer Society, 2009:254-258.
[9] Chuang L Y, Yang C H, Wu K C, et al. A hybrid feature selection method for DNA microarray data. Computers in Biology & Medicine, 2011, 41(4):228-237.
[10] Shen Q, Shi W M, Kong W, et al. A combination of modified particle swarm optimization algorithm and support vector machine for gene selection and tumor classification. Talanta, 2007, 71(4):1679-1683.
[11] Joroughi M, Shamsi M, Saberkari H, et al. Gene selection and cancer classification based on microarray data using combined BPSO and BLDA algorithm. Journal of Thoracic & Cardiovascular Surgery, 2014, 5(2):1931-9.
[12] Han, F., Yang, C., Wu, Y.Q, et al. A gene selection method for microarray data based on binary pso encoding gene-to-class sensitivity information. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2017, 14(1), 85-96.
[13] Han F, Sun W, Ling Q H. A novel strategy for gene selection of microarray data based on gene-to-class sensitivity information. Plos One, 2014, 9(5):e97530.
[14] Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: a new learning scheme of feedforward neural networks. In: IEEE International Joint Conference on Neural Networks, 2004, pp. 985-990.
[15] Yuan Y, Ji X, Sun Z, et al. Application of Isomap for cluster analyses of gene expression data. Journal of Tsinghua University, 2004, 44(9):1286-1289.