Advances in Biological Sciences Research, volume 3

International Conference on Biological Engineering and Pharmacy (BEP 2016)

A Novel Feature Selection Method for Gene Expression Data Based on Samples Localization

Mingyue SHENG
College of Computer Science and Technology, Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education
Jilin University
Changchun, China
e-mail: smy0917@foxmail.com

Wei DU*
College of Computer Science and Technology, Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education
Jilin University
Changchun, China
e-mail: weidu@jlu.edu.cn

Yuan TIAN
College of Computer Science and Technology, Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education
Jilin University
Changchun, China
e-mail: yuant2012@163.com

Yanchun LIANG*
College of Computer Science and Technology, Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education; Zhuhai Laboratory of Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education
Jilin University; Zhuhai College of Jilin University
Changchun, China; Zhuhai, China
e-mail: ycliang@jlu.edu.cn

***Abstract*—It is an important and hot topic for researchers to develop an efficient and robust feature selection method from gene expression profile data with thousands of genes and small sample size. At present, most of feature selection methods are constructed models to use all samples of gene expression data, but these methods are never considered the influence of outlier samples and the distribution of samples. Besides, it is well known that cancer is a kind of heterogeneous disease, and different cancer tissue samples of same organs have many different subtypes on molecular characteristics. So, we should select samples with the same genetic characteristics to construct models. Therefore, in this article, we proposed a novel and efficient feature selection approach based on localized samples to extract gene signatures more accurately. We picked out the nearest samples in a certain range for each target sample and obtained the best localized samples by constructing a sample-sample similarity network, which calculated Euclidean distance between the central samples with others by using gene expression values firstly. Secondly, we established the co-expression networks by selecting top nearest samples, and formed steady-state probability network applying to Random Walk with Restart (RWR) method. Finally, through dividing into this network and comparing five selection strategies, we got localized samples for best cancer classification. We applied our method on six datasets across different cancer types. The average accuracies of top 100 genes of the method by SVM classifiers in leave-one-out cross validation (LOOCV) are 95.46%, 94.01%, 96.20%, 99.79%, 99.08% and 99.37%, respectively. The results show that the proposed method obtains excellent performance on these datasets. It also indicates that the proposed method is effective and applicable.**

Keywords-feature selection; samples localization; cancer classification

I. INTRODUCTION

The development of high-throughput gene chip technique has resulted in generation of a large number of gene expression profile data measured some mRNA experiments by researchers. The microarray data can obtain from the public database including GEO [1] (Gene Expression Omnibus) and SMD [2] (Stanford Microarray Database). Recently, with the advent of next-generation sequencing (NGS) technology, it is possible to use RNA sequencing [3] (RNA-seq) data, which can download from TCGA [4] (The Cancer Genome Atlas) database, to deal with the research of gene expression on high-throughput genomics. Since the characteristic of gene expression profiling data with tens of thousands of genes and a small number of samples may hinder extracting useful information, it is extremely necessary to propose an efficient and robust feature selection method to extract information genes from gene expression data. In the last two decades, feature selection technique has already been a significant and efficient tool for analyzing and applying on gene expression data. These methods find relevant subset of features and eliminate irrelevant, redundant features, in order to tackle high-dimension cancer data with a large number of genes and small samples size in the field of bioinformatics [5] [6].

However, it is well known that cancer is a kind of heterogeneous disease, and different cancer tissue samples of same organs have many molecular characteristics on different subtypes [7]. In addition, patients with the same genetic characteristics might share same molecular mechanisms during the development and evolution of cancer [8]. So, using localized samples with the same genetic characteristics to train more accurate models and predict whether one person has cancer will be more important. A new problem which we need to face is how to pick out more similar samples and how to improve the classification evaluation. In the past, there were two common ways to select features, the one way was using all samples to select features [5] [6] [9], the other was using a part of all samples randomly [10] [11] for selecting samples. However, selecting features by all samples ignored outlier samples and the distribution of samples, and selecting features by samples randomly were blind and baseless to ignore the characteristic of data and improved the rate of prediction error [12].

In this paper, we propose an efficient and easily useful feature selection method based on localized samples. Localized samples are a subset of samples from original dataset, and the method using them to train feature selection models can get better classification performance because the method can reduce the influence of outlier samples and distribution of samples. We only picked out the nearest samples compared to each target sample in a certain range to train the models. The best localized samples were gotten by establishing and dividing into the similarity network between any two samples, and there are three steps to establish the similarity network. Firstly, we calculated Euclidean distance between the central samples with their neighbor samples by using gene expression values. Secondly, we established the co-expression networks by selecting top four nearest samples. Thirdly, according to Random Walk with Restart (RWR) method, we formed the steady-state probability network. And then, we divided into this network by different threshold values and compared these selection strategies, and obtained the localized samples for best cancer classification. We evaluated the proposed method by using leave-one-out cross validation (LOOCV) on breast, gastric, pancreatic, lung, thyroid and prostate cancer from GEO and TCGA database. The best accuracies of the proposed method on these datasets for top 100 genes by SVM classifiers were 98.51%, 97.27%, 98.55%, 100%, 100% and 100%, respectively. And, the average of the method on six datasets were 95.46%, 94.01%, 96.20%, 99.79%, 99.08% and 99.37%, respectively. The results showed that the proposed method based on localized samples is superior to other methods in best and mean accuracies.

II. MATERIAL AND METHOD

A. Data Description

We apply six cancer types of gene expression profiling data to measure the performance of different feature selection methods, four of which are downloaded from GEO database [1], and others are downloaded from TCGA database [4]. The detail information of six datasets including breast, gastric, lung, pancreatic, thyroid and prostate cancer are summarized in Table I.

TABLE I. THE GENE EXPRESSION DATASETS OF SIX CANCER TYPES

Database	Datasets Information			
	CancerType	*Datasets*	*#SampleNum (+/-)*	*GeneNum*
GEO	Breast	GSE15852	86(43/43)	12591
	Gastric	GSE27342	160(80/80)	16463
	Lung	GSE19804	120(60/60)	16894
	Pancreatic	GSE28735	90(45/45)	20254
TCGA	Thyroid	THCA	112(56/56)	20500
	Prostate	PRAD	86(43/43)	20500

B. Feature Selection Based on Samples Localization

In some ways, localizing samples not only can represent all samples, but can get better performance for selecting features than using all samples. During the development and evolution of cancer, the samples with the same genetic characteristics may share same molecular mechanisms [8]. At the same time, compared with training by using all samples, training by using part of similar samples can get better performance. Hence, we only pick out the most similar samples for each sample to select features and train classifiers.

First of all, we define two concepts in this paper. The first concept is called central sample. There is an opportunity for each sample in original dataset to be the central sample. Then, we regard each central sample as a target sample, and calculate similarity between the target samples and other samples. According to the defined thresholds, we can get the most similar samples for each central sample, which can be used to train the classifier having high accurate. The second concept is called mini-classifier. We can construct sets of classifiers just using the most similar samples of these central samples. The sets of classifiers are regarded as mini-classifiers. Then, we proposed the feature selection method based on samples localization, which also can be called SL (Sample Localization) feature selection for short. There are four main steps of the proposed method and the framework is shown in Fig. 1.

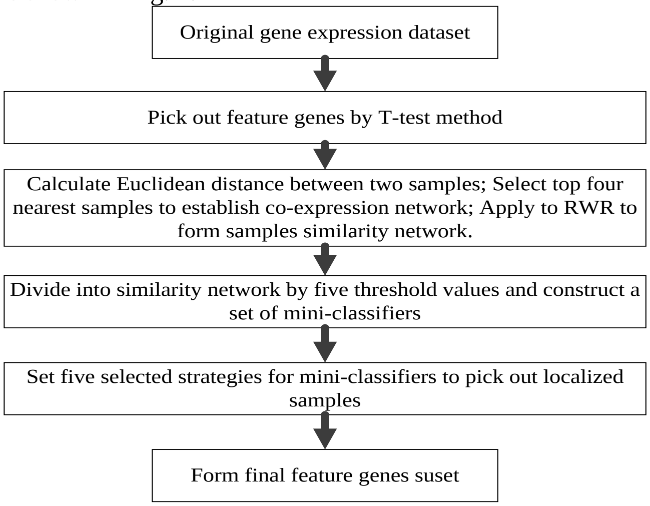

Original gene expression dataset
Pick out feature genes by T-test method
Calculate Euclidean distance between two samples; Select top four nearest samples to establish co-expression network; Apply to RWR to form samples similarity network.
Divide into similarity network by five threshold values and construct a set of mini-classifiers
Set five selected strategies for mini-classifiers to pick out localized samples
Form final feature genes suset

Figure 1. The framework of proposed feature selection methods based on samples localization

Firstly, we calculate Euclidean distance between each central sample with other neighbor samples by using their gene expression values, and then we establish the co-expression networks [13] by selecting top four nearest samples of the central sample. The Euclidean distance is obtained according to the following formula:

$$d(i,j)=\left\|S_i-S_j\right\|=\sqrt{\sum_{k=1}^{g}(S_{ik}-S_{jk})^2} \tag{1}$$

where $d(i,j)$ denotes the Euclidean distance of sample i and sample j, $S_i(S_{i1},S_{i2},S_{i3},...,S_{ig})$ and $S_j(S_{j1},S_{j2},S_{j3},...,S_{jg})$ denote the vector of gene expression values on these two samples, respectively. Here, g denotes the genes number. And, we apply the random walk with restart (RWR) method [14] [15] to form steady-state probability network of all central samples. Before using the RWR method, we normalize the adjacency matrix of co-expression networks by column, which makes each column sum to 1. And this method is obtained according to the following formula:

$$P_{ij}=(1-c)AP_{i(j\text{-}1)}+cV \tag{2}$$

where P_{ij} denotes the steady-state probability of central sample i randomly walking to each sample j on the network, and $P_{i(j\text{-}1)}$ denotes the steady-state probability of central sample i randomly walking to each sample $(j\text{-}1)$ on the network. A denotes the column-normalized adjacency matrix, V denotes an identity matrix, and c denotes restarting probability, which is a fixed value to return to the sample i . We aren't able to get steady-state probability matrix until P_{ij} has been converged. Each probability is defined as the similarity score between two central samples, and the bigger probability has the higher similarity. Finally, we can get the samples similarity network.

Secondly, we set five defined thresholds to divide each samples similarity network according to the similarity of samples. The purpose of dividing the network is to pick out more similar samples of center samples. The five thresholds are 0.00005, 0.0001, 0.0002, 0.0005, and 0.001, respectively. If the similarity between one central sample and one of its neighbor samples is higher than the threshold value, the neighbor sample is selected, and otherwise, the neighbor sample is not selected.

Then, after dividing the network, we obtain five different subsets of similar samples for each central sample according to five different thresholds. After that, in order to verify the performance of mini-classifiers, we select each central sample as testing data and select five subsets of samples of central samples as training data, respectively. We construct and test mini-classifiers using training and testing data by support vector machine (SVM) [16] method. It is noticed that the subsets of samples do not include its central sample.

Finally, we also set five different selection strategies to pick out localized samples according the number of correct mini-classifiers of each central sample. Based on experimental verification and analysis, we obtained the optimal selection strategy from five strategies and got a set of central samples of original dataset according to the optimal strategy. The set of central samples is the localized samples which we need.

C. Classification on Samples Localization

In this article, we apply T-test as the feature selection method to select final informative gene lists by localized samples. And then, SVM classifier [16] is applied to train classification models and leave-one-out cross validation method (LOOCV) is used to evaluate the performance of classification. We pick out top 100 significant genes as the features list according to *p* value of ascending by the T-test. In order to ensure the integrity of validating process, it is necessary to divide original dataset into two parts before feature selection. We need to select one sample each time by LOOCV as testing data and the rest of samples as training data until all samples have been completed.

III. Results

A. Assessment Measurement

In this paper, we use LOOCV method to evaluate the performance of various methods on different datasets. There are two different measurements to evaluate performance of different feature selection method: optimal performance and average performance [17]. The optimal performance is the best classification result with different number of selected feature genes, and the average performance is the mean performance with different number of selected feature genes. In this article, we use three criterions, *accuracy* (*ACC*), *sensitivity* (*SEN*) and *specificity* (*SPE*), to measure the classification. They represent the rate of correct classification results, the rate of patients who are correctly identified and the rate of healthy people who are correctly identified, respectively [18]. The three criterions are obtained according to the following formula:

$$\begin{aligned}ACC&=\frac{TP+TN}{TP+FP+TN+FN}\\ SEN&=\frac{TP}{TP+FN}\\ SPE&=\frac{TN}{TN+FP}\end{aligned} \tag{3}$$

where TP denotes true positive, the number of cancer samples correctly identified as cancer, FP denotes false positive, the number of normal samples incorrectly identified as cancer, TN denotes true negative, the number of normal samples correctly identified as normal, FN denotes false negative, the number of cancer samples incorrectly identified as normal.

B. *Performance of Different Selection Strategies*

In order to compare the performance of five selection strategies, we apply them on four types of cancer datasets containing gastric, pancreatic, thyroid and prostate cancer. The accuracies of different selection strategies are shown in Fig. 2 for gastric and pancreatic cancer and shown in Fig. 3 for thyroid and prostate cancer. In each figure, there are five curves, and each curve is represented one selection strategy. As we can see, in gastric cancer, the highest classification accuracy is 97.27% when used number of correct mini-classifiers is greater than or equal to 5. For pancreatic cancer, the highest classification accuracy is 98.55% when the used number of correct mini-classifiers is greater than or equal to 5, while the highest accuracy of other four selection strategies is 94.32%. Through calculating and analyzing five selection strategies, we can discover that the classification performance of localized samples is the best when the used number of correct mini-classifiers is greater than or equal to 5.

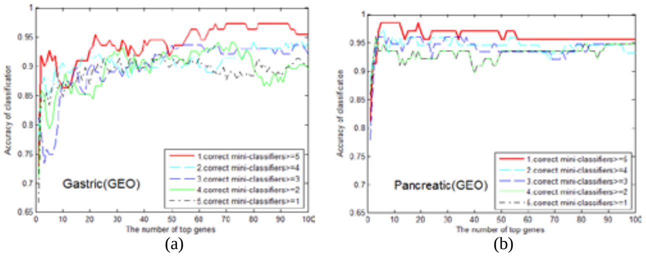


Figure 2. The accuracies of different selection strategies for gastric and pancreatic cancer. Five different color curves of each figure represent the selection strategies of different cancer datasets, respectively. (a)Gastric cancer. (b)Pancreatic cancer

As for the best selection strategy, the number of eliminated samples is accounted for 3.63%, 5.95% of the total samples on thyroid and prostate cancer downloaded from TCGA respectively, while the number of eliminated samples of gastric and pancreatic cancer of GEO is accounted for 30.38%, 21.59% respectively. Furthermore, the best accuracy of best selection strategy averagely increases by 0.93%, 1.22% on thyroid, prostate cancer compared with other selection strategies respectively, while gastric and pancreatic cancer of GEO averagely increased by 3.79%, 2.12%, respectively. We can infer that the batch effect of GEO data is much higher than TCGA data.

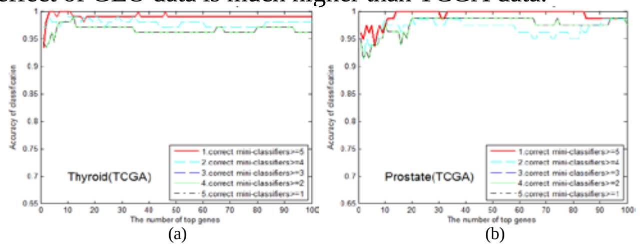


Figure 3. The accuracies of different selection strategies for thyroid and prostate cancer from TCGA. Five different color curves of each figure represent the selection strategies of different cancer datasets, respectively. (a) Thyroid cancer. (b) Prostate cancer

C. *Performance Evaluation on Different Methods*

We compare SL feature selection method with T-test, RankSum (the rank sum test method), mRMR [19] (minimum redundancy-maximum relevance filter method), and Random Sample method on multiple datasets. As for random sample methods, we randomly choose 80% of normal samples and 80% cancer samples as final experimental datasets [10]. We have gotten best accuracy of six datasets of Breast, Gastric, Pancreatic, Lung, Thyroid and Prostate cancer applying five different methods, which is illustrated in Table II. The number in the bracket represents the number of top genes of reaching best accuracies. For example, on pancreatic dataset, the top 5 selected genes by SL feature selection method make the highest accuracy up to 98.55%. At the same time, the mean accuracy of six datasets is shown in Table III. The best accuracies of six cancer datasets of our method for top 100 selected genes are 98.51%, 97.27%, 98.55%, 100%, 100% and 100%, respectively. The mean accuracies of six cancer datasets of our method for top 100 selected genes are 95.46%, 94.01%, 96.20%, 99.79%, 99.08% and 99.37%, respectively. From two tables, SL feature selection method performs perfectly better than other four methods both in best accuracy and average accuracy. The performance of our method on the various datasets has illustrated its effectiveness in feature selection. Though calculating, the average accuracies of localized samples on six cancer datasets which is higher than all samples' are 11.4%, 15.78%, 11.4%, 7.81%, 4.22%, and 5.64%, respectively. It can indicate that our method using localized samples has much better performance than using all samples.

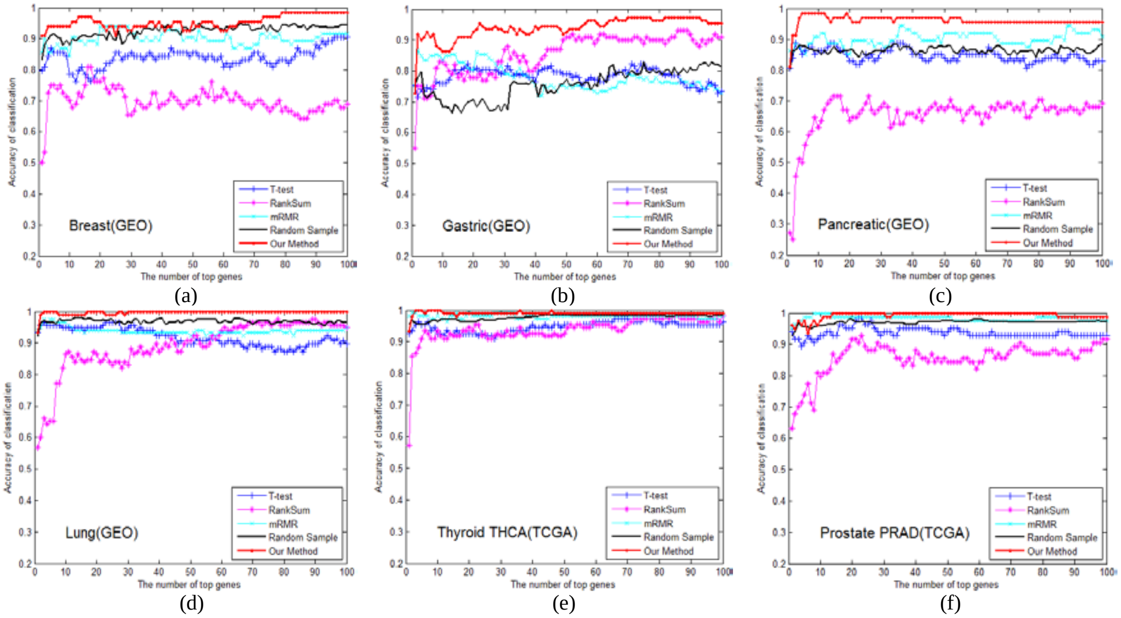
TABLE II. THE BEST ACCURACY OF FIVE METHODS ON SIX DATASETS (%)

Datasets	**Feature Selection Methods**				
	T-test	***RankSum***	***mRMR***	***Random Sample***	***SL Feature Selection***
Breast	90.48(96)	80.95(16)	94.05(20)	95.29(52)	**98.51(79)**
Gastric	82.28(18)	93.00(86)	86.71(2)	83.02(97)	**97.27(66)**
Pancreatic	88.64(3)	71.59(15)	94.32(36)	88.57(34)	**98.55(5)**
Lung	96.61(3)	97.46(78)	97.46(3)	98.30(13)	**100(3)**
Thyroid	97.27(66)	98.18(77)	99.09(2)	98.41(44)	**100(3)**
Prostate	98.81(22)	92.86(23)	**100(8)**	98.24(19)	100(14)
Mean	92.35	89.01	95.27	93.75	98.90

TABLE III. THE AVERAGE ACCURACY OF FIVE METHODS ON SIX DATASETS (%)

Datasets	Feature Selection Methods				
	T-test	***RankSum***	***mRMR***	***Random Sample***	***SL Feature Selection***
Breast	84.06	70.06	90.13	92.54	**95.46**
Gastric	78.23	85.67	78.11	75.65	**94.01**
Pancreatic	84.80	65.45	90.19	86.49	**96.20**
Lung	91.98	89.19	93.88	96.94	**99.79**
Thyroid	94.86	93.89	98.07	97.67	**99.08**
Prostate	93.73	85.56	98.36	97.02	**99.37**
Mean	87.94	81.64	91.46	91.05	**97.32**

In addition, the prediction results of top 100 genes by five methods are shown in Fig. 4. In the figure, it is shown that the accuracies of SL feature selection method are better than other methods on most of datasets, especially the four datasets from GEO database. For example: In breast dataset, the top 79 selected genes can reach the best classification accuracy as 98.51%, which is higher than other methods. In pancreatic dataset, the top 5 selected genes can reach the best classification accuracy as 98.55%, which is higher than others methods. The mRMR method performs well on thyroid and prostate cancer from TCGA, especially on prostate cancer. The mRMR method gets the best performance on prostate cancer, and the top 8 selected genes can reach best accuracy as 100%. However, using SL feature selection method, the top 14 selected genes get the accuracy as 100%. Although the result of SL feature selection method on prostate cancer is not the best, it is very close to the best value.



(a) (b) (c) (d) (e) (f)

Figure 4. Top 100 gene prediction results by five methods on six cancer datasets. The figure represents the classification accuracy of six cancer type, respectively. (a): Breast cancer. (b): Gastric cancer. (c): Pancreatic cancer. (d): Lung cancer. (e): Thyroid cancer. (f): Prostate cancer. Different color curves of each cancer figure particularly represent five diverse methods, which include T-test, RankSum, mRMR, Random Sample and SL feature selection method. Among the figure, X axis represents the number of selected genes and Y axis represents the classification accuracies.

D. Performance on Validation Datasets

In order to ensure the performance of the best model, we need to validate various datasets of same cancers by the model. Here, we take lung cancer as an example. Firstly, the best classification model of lung cancer has been obtained by SL feature selection method and SVM. Then, we apply the model to five microarray datasets of lung cancer from GEO database to validate model by accuracy, specificity and sensitivity. The results are listed in Table IV. Based on the results, we find that our proposed method can obtain satisfactory performance of classification on other datasets of same cancer type. The method can get satisfactory performance on the five groups of lung cancer datasets, and the accuracy, specificity and sensitivity on microarray series GSE18842 are 97.73%, 100%, 95.65%, respectively. The worst performance is on microarray series GSE31552, which accuracy, specificity and sensitivity are 73.40%, 86.67% and 67.19%.

TABLE IV. THE VALIDATION RESULTS OF FIVE LUNG CANCER DATASETS

#Series	Validation Results			
	Sample/Gene	***Accuracy***	***Specificity***	***Sensitivity***
GSE18842	88/21889	97.73%	100%	95.65%
GSE7670	54/11948	94.44%	90%	100%
GSE2514	38/8980	86.84%	85%	88.89%
GSE10072	66/12591	83.33%	100%	75%
GSE31552	94/17945	73.40%	86.67%	67.19%

IV. Conclusion and Discussion

In recent years, feature selection technology has already become a crucial tool for analysis gene expression data and most of these methods obtain models by using all samples of original data. However, cancer as a kind of heterogeneous disease, the different samples with tissues of same organs may have many different subtypes and only these samples with same genetic characteristic may share same molecular mechanisms. Therefore, in this article, we propose an efficient, robust and easily useful feature selection method based on localized samples to pick out gene markers, which can get satisfactory performance by using a subset of all samples. Through calculating Euclidean distance, establishing sample co-expression network and forming steady-state probability network by random walk with restart (RWR) method, we construct a sample-sample similarity network to select localized samples, which can get the better classification result than using all samples.

We apply the proposed SL feature selection method on six different cancer datasets, including breast, gastric, lung, pancreatic, thyroid and prostate cancer. In the aspect of performance evaluation, we focus on accuracy, sensitivity and specificity as our measurement standard. The best performance and average performance on six datasets by our method are (98.51%, 95.46%), (97.27%, 94.01%), (100%, 99.79%), (98.55%, 96.20%), (100%, 99.08%) and (100%, 99.37%), respectively. We compare the result of localized samples to all samples; the result of localized samples is higher than all samples' on six cancer datasets. In addition, comparison with other four methods, T-test, RankSum, mRMR and random samples method, the mean of best accuracy and the mean of average accuracy on all datasets are both the best results. It is shown that SL feature selection method has an excellent classification performance compared with some other existing methods, which shows potential applications and capabilities on analysis of gene expression data.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (61272207, 61472158, 61402194, and 61572228) and China Postdoctoral Science Foundation 2014T70291.

References

[1] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, et al., "NCBI GEO: archive for high-throughput functional genomic data," Nucleic Acids Research, vol. 37, 2009, pp. D885-D890.

[2] J. Hubble1, J. Demeter, H. Jin, M. Mao, M. Nitzberg, T. B. K. Reddy, et al., "Implementation of GenePattern within the Stanford Microarray Database," Nucleic Acids Research, vol. 37, 2009, pp. D898-D901.

[3] A. K. Yim, J. W. Wong, Y. S. Ku, H. Qin, T. F. Chan, H. M. Lam, "Using RNA-Seq Data to Evaluate Reference Genes Suitable for Gene Expression Studies in Soybean," Plos One, vol. 10, Sept. 2015, doi: 10.1371/journal.pone.0136343.

[4] C. G. A. Network, "Comprehensive molecular portraits of human breast tumours," Nature, vol. 490, Oct. 2012, pp. 61-70, doi:10.1038/nature11412.

[5] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J.M. Benítez, F. Herrera, "A review of microarray datasets and applied feature selection methods," Information Sciences, vol. 282, 2014, pp. 111-135.

[6] R. K. Singh, M. Sivabalakrishnan, "Feature Selection of Gene Expression Data for Cancer Classification: A Review," Procedia Computer Science, vol. 50, 2015, pp. 52-57, doi: 10.1016/j.procs.2015.04.060.

[7] G. Bianchini, T. Iwamoto, Y. Qi, C. Coutant, C. Y. Shiang, B. Wang, et al., "Prognostic and therapeutic implications of distinct kinase expression patterns in different subtypes of breast cancer," Cancer Research, vol. 70, Nov. 2010, pp. 8852-8862.

[8] M. J. Jahid, T. H. Huang, J. Ruan, "A personalized committee classification approach to improving prediction of breast cancer metastasis," Bioinformatics, vol. 30, Mar. 2014, pp. 1858-66.

[9] G. D. Zhao, Y. Wu, "Feature Subset Selection for Cancer Classifiation Using Weight Local Modularity," Scientific Reports, vol. 6, Oct. 2016, pp. 1-16, doi: 10.1038/srep34759.

[10] L. H. He, Z. B. Cao, Y. Wang, W. D, Y. C. Liang, "An Ensemble Feature Selection Method Based on mRMR for Paired Microarray Data," Journal of Computation Information Systems, June 2014, pp. 4875-4882, doi: 10.12733/jcis10628.

[11] A. R. Shafay, N. Balakrishnan, J. Ahmadi, "Bayesian prediction of order statistics with fixed and random sample sizes based on k-record values from Pareto distribution," Communication in Statistics-Theory and Methods, Feb. 2016, pp. 721-735, doi: 10.1080/03610926.2015.1004093.

[12] L. R. van Hoeven, M. P. Janssen, K. C. B. Roes, H. Koffijberg, "Aiming for a representative sample: Simulating random versus purposive strategies for hospital selection," BMC Medical Research Methodology, vol. 15, 2015, pp. 1-9, doi: 10.1186/s12874-015-0089-8.

[13] J. H. Ruan, A. K. Dean, W. Zhang, "A general co-expression network-based approach to gene expression analysis: comparison and applications," Bmc Systems Biology, vol. 4, Feb. 2010, pp. 1-21, doi: 10.1186/1752-0509-4-8.

[14] C. Zhang, S.Jiang, Y. C. Chen, Y. D. Sun, J. W. Han, "Fast Inbound Top-K Query for Random Walk with Restart," Mach Learn Knowl Discov Databases, Sep. 2015, pp. 608-624, doi: 10.1007/978-3-319-23525-7_37.

[15] H. H. Tong, C. Faloutsos, J. Y. Pan, "Fast Random Walk with Restart and Its Applications," Proceedings of the Sixth International Conference on Data Mining, IEEE Computer Society, 2006, pp. 613-622.

[16] C. Cortes, V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, 1995, pp. 273-297.

[17] Y. Tang, Y. Q. Tang, Z. Huang, "Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 4, July 2007, pp. 365-381, doi: 10.1109/TCBB.2007.70224.

[18] M. Majnik, Z. Bosnić, "ROC analysis of classifiers in machine learning: A survey," Intelligent Data Analysis, vol. 17, 2013, pp. 531-558, doi: 10.3233/IDA-130592.

[19] C. Ding, H. C. Peng, "Minimum redundancy feature selection from microarray gene expression data," Journal of Bioinformatics & Computational Biology, vol. 3, 2005, pp. 185-205.