Osong Public Health Res Perspect 2014 5(5), 279—285
http://dx.doi.org/10.1016/j.phrp.2014.08.004
pISSN 2210-9099 eISSN 2233-6052

ORIGINAL ARTICLE

A New Direction of Cancer Classification: Positive Effect of Low-Ranking MicroRNAs

Feifei Li, Minghao Piao, Yongjun Piao, Meijing Li, Keun Ho Ryu*

Database and Bioinformatics Laboratory, College of Electrical and Computer Engineering, Chungbuk National University, Cheongju, Korea.

Received: July 20, 2014
Revised: August 15, 2014
Accepted: August 16, 2014

KEYWORDS:
cancer classification,
feature selection,
low-ranking miRNAs,
miRNA expression profile

Abstract

Objectives: Many studies based on microRNA (miRNA) expression profiles showed a new aspect of cancer classification. Because one characteristic of miRNA expression data is the high dimensionality, feature selection methods have been used to facilitate dimensionality reduction. The feature selection methods have one shortcoming thus far: they just consider the problem of where feature to class is 1:1 or n:1. However, because one miRNA may influence more than one type of cancer, human miRNA is considered to be ranked low in traditional feature selection methods and are removed most of the time. In view of the limitation of the miRNA number, low-ranking miRNAs are also important to cancer classification.
Methods: We considered both high- and low-ranking features to cover all problems (1:1, n:1, 1:n, and m:n) in cancer classification. First, we used the correlation-based feature selection method to select the high-ranking miRNAs, and chose the support vector machine, Bayes network, decision tree, k-nearest-neighbor, and logistic classifier to construct cancer classification. Then, we chose Chi-square test, information gain, gain ratio, and Pearson's correlation feature selection methods to build the m:n feature subset, and used the selected miRNAs to determine cancer classification.
Results: The low-ranking miRNA expression profiles achieved higher classification accuracy compared with just using high-ranking miRNAs in traditional feature selection methods.
Conclusion: Our results demonstrate that the m:n feature subset made a positive impression of low-ranking miRNAs in cancer classification.

1. Introduction

Chronic lymphocytic leukemia [1] is the first known human disease that is associated with microRNA (miRNA) deregulation. Many miRNAs have been found to have a connection with some types of human cancer [2,3]. Thus, a great deal of research has been done regarding machine learning methods to analyze cancer classification using miRNA expression profiles. From the year 1993, when the first identified miRNA [4] was discovered until now, only thousands of miRNAs have been discovered. The limitation of sample availability

*Corresponding author.
E-mail: khryu@dblab.chungbuk.ac.kr

leads to the high dimensionality [5] of miRNA expression data. The high dimensionality may cause a series of problems for cancer classification, such as added noise, reduced accuracy rate, and increased complexity. Although both feature selection and feature extraction can be used to reduce dimensionality, feature selection is a better choice than feature extraction for miRNA expression data. Feature selection is used in areas where there are a large number of features compared with the small number of samples, which is a characteristic of miRNA expression data; the goal of feature extraction is to create new features using some transform functions of the original features, but these new features cannot be explained in the physical aspect.

Lu et al [6] used a new bead-based flow cytometric miRNA expression profiling method to analyze 217 mammalian miRNAs from 334 samples. The k-nearest-neighbor (KNN) classification method was used to classify the normal and tumor samples, and the probabilistic neural network (PNN) algorithm was adopted to perform the multi-class predictions of poorly differentiated tumors. The results showed the potential of miRNA profiling in cancer diagnosis. Based on this study, many further researches have been done using different machine learning methods. In Zheng and Chee's work [7], the discrete function learning (DFL) algorithm was used for the miRNA expression profiles to find the subset of miRNAs. The selected miRNAs were used to classify normal and tumor samples, and at last they find some important miRNAs for normal/tumor classification. Xu et al [8] used particle swarm optimization (PSO) for miRNA selection, and default adaptive resonance theory (ART) neural network architectures (ARTMAP) to classify multiple human cancers. The results showed that cancer classification can be improved with feature selection. Kim and Cho [9] adopted seven feature selection methods to reduce dimensionality of miRNA expression data and built binary class classification. They draw the conclusion that the proper combination of feature selection and classification method is important for cancer classification.

Thus far the feature selection methods attempt to rank features based on some evaluation metric and select the high-ranking features. These high-ranking features indicate the relationship between feature and class is 1:n and n:1, which means these features can produce pure class. However, the miRNA expression data are different from others in that one miRNA may have influence for more than one type of cancer [10], like the microRNA-21, which is related to both glioblastoma and astrocytoma. However, these miRNAs are considered as low-ranking features and removed during feature selection. Because of the limitation of the miRNA number, it is reasonable to take this type of miRNA into consideration during cancer classification. Therefore, in our study, we made a new hypothesis that considers both the high- and low-ranking features covers all the cases (1:1, n:1, 1:n, and m:n) and can provide better accuracy in cancer classification. We used the data resource from the work of Lu et al [6], and adopted different types of feature selection methods with different classifiers to do the analysis. Finally, the results proved that the m:n features can lead to higher classification accuracy compared with the traditional feature selection methods, and it is reasonable to take the low-ranking features into consideration for cancer classification.

2. Materials and methods

The goal of feature selection is to remove the redundant and irrelevant features to find a subset of features. Feature selection involves two aspects: evaluation of a candidate feature subset using some evaluation criterion, and searching through the feature space to select a minimum subset of features. The categories of feature selection algorithms can be identified based on their evaluation metrics: wrapper, filter, and embedded methods. Filter methods first calculate the relevance score for each feature, then rank each feature according to some univariate metric, and then select the high-ranking features. The univariate metric of most proposed techniques means each feature is considered separately, thus ignoring feature dependencies. However, the multivariate filter methods are geared toward the incorporation of feature dependencies. One typical multivariate filter method is the correlation-based feature selection (CFS) [11]. It ranks feature subsets according to a correlation-based heuristic evaluation function which is biased toward subsets that contain features that are highly correlated with the class and uncorrelated with each other.

Because there is no evidence to show which type of feature selection method would fit for miRNA expression data, we chose many different methods for the analysis and compared their results. First, we used the CFS with different search algorithms. Then, we used the ranker search method with different attribute evaluators. The information regarding these methods is shown in Table 1.

Table 1. Information on feature selection method.

Attribute evaluator	Search method
Correlation-based feature selector	Re-ranking
	Best first
	Particle swarm optimization
	Tabu
Pearson's correlation	Ranker search
Chi-square	
Information gain	
Gain ratio	

3. Results

3.1. Data set

The miRNA expression data used in this paper are from the work of Lu et al [6]. The data are used to build a multiclass classifier and consist of five types of tumor samples from the colon, uterus, pancreas, T cell acute lymphoblastic leukemia (ALL), and B cell ALL which include 73 samples with the expression value of 217 miRNAs for multiple cancer types. Details regarding the cancer types are shown in Table 2.

3.2. Performance evaluation

To obtain a reliable result, 10-fold cross validation is performed on the entire data set. The data set is randomly divided into 10 parts; nine of them are used as a training set, the 10th as part of a test set.

3.3. Analysis of high-ranking miRNAs

In our study, we first used CFS with four different search methods including re-ranking search, best first search, tabu search, and PSO search method to reduce the dimensionality. For comparison, we tested these selected features on five classifiers including LibSVM algorithm [12] of the support vector machine (SVM) classifier [13], the Bayes network classifier [14], C4.5 algorithm of the decision tree classifier [15], the KNN classifier, and the logistic classifier [16].

The result is shown in Figure 1. We first tested these classification methods without feature selection, and saw that the SVM obtained the best result. Then we reduced the dimensionality with CFS, using those search methods that can automatically select the features with the exact number. The re-ranking search method resulted in 15 top-ranking features, the best search method resulted in 16 top-ranking features, the tabu search method resulted in 17 top-ranking methods, and the PSO search method resulted in 50 top-ranking features. For SVM classifier, these feature selections cannot improve the accuracy compared with the accuracy without feature selection. For the Bayes network classifier, the accuracy was improved when using re-ranking, best first, and tabu search methods. The PSO search method did not show good results for the Bayes network classifier, but showed a good result for the decision tree and KNN classifier. For logistic classifier, accuracy was improved when using best first and tabu search methods, and the accuracy was highest in all of these results.

Table 2. The number of the samples for each cancer type.

Cancer type	Number of tumor samples
Colon	10
Pancreas	9
Uterus	10
B cell ALL	26
T cell ALL	18
Total	73

The results indicated that feature selection is necessary for cancer classification. Because these methods just selected the fit number of features, it is difficult to determine how the number of features influences the classification accuracy. Therefore, we performed another experiment using Pearson's correlation, Chi-square distribution, information gain, and gain ratio as the attribute evaluator to determine the relationship between the number of features and classification accuracy. In addition, these five classification methods (SVM, Bayes network, decision tree, KNN, and logistic) were adopted to build the classifier. Figures 2–5 show the classification accuracy using these four different feature selection methods. The number of the top-ranking features chosen for testing is from 10 to 200.

Comparing the results of the four feature selection methods, the Pearson's correlation method and gain ratio method show similar results, whereas the Chi-square method and the information gain method show similar results. When the number of features is very small, the accuracy of Pearson's correlation method and the gain ratio method is very low, but the accuracy of the Chi-square method and the information gain method is high. For all of these feature selection methods, there is a similar trend that with the increase of the feature numbers the accuracy is also improved.

Compared with different classification methods, the Bayes network classifier showed the worst accuracy for all of the four feature selection methods; by contrast, the SVM classifier showed a relative advantage in the cancer classification using miRNA expression data. Also, for decision tree classifier, when using Chi-square statistic and information gain feature selection methods, higher accuracy can be achieved when the feature number is small compared with other classification methods.

3.4. Analysis of low-ranking miRNAs

Both of these feature selection methods select the high-ranking features, but as we mentioned previously, some low-ranking miRNAs are also very important to cancer classification. Therefore, we considered both the high- and low-ranking features to form the m:n feature subset. The previous experiment shows that the SVM classifier showed the better results, and we compared the accuracy of these four feature selection methods with SVM classifier in Figure 6. It shows that the information gain and Chi-square feature selection methods are better compared with the other two methods. When the feature number is <30, Pearson's correlation and gain ratio feature selection methods show very low classification accuracy, which means these selected top-ranking features cannot accurately classify the miRNA data. Considering this reason, the information gain and Chi-square feature selection methods were used to form the feature subsets with both

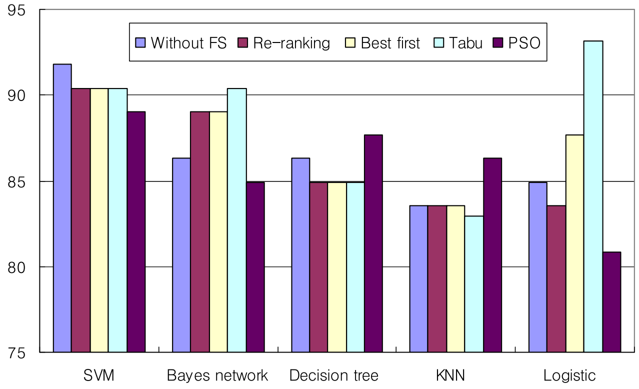

Figure 1. Classification accuracy (%) of CFS for high-ranking features. KNN = k-nearest-neighbor; PSO, particle swarm optimization; SVM, support vector machine.

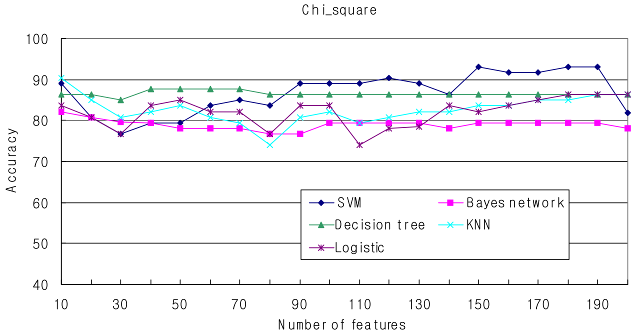

Figure 2. Classification accuracy (%) with Chi-square feature selection. KNN = k-nearest-neighbor; SVM, support vector machine.

high- and low-ranking features, and the LibSVM package of SVM classifier was selected for the multiple classification problems.

The results are shown in Table 3. First we selected 10 high-ranking features, which means the relationship between feature and Class is 1:1 or n:1. The information on selected high-ranking miRNA is shown in Tables 4 and 5. The classification accuracy is 89.04% for both the information gain and Chi-square statistic feature selection methods. Next, we considered the case of the feature to class is 1:n; in this instance we selected 17 low-ranking features. The information on selected low-ranking miRNA is shown in Tables 6 and 7. The classification accuracy of the information gain method is 52.05%, whereas the classification accuracy of the Chi-square method is 50.68%. Obviously the accuracy is very low because the low-ranking features would lead to the impurity of the class. Last, we considered the m:n features with both the high- and low-ranking features, and in this condition feature to class is m:n. We combined both the 10 high-ranking features and 17 low-ranking features, for a total of 27 features, and used them to assign the classification; surprisingly, a very good result was achieved. The classification accuracy of information gain method is 94.52% and the classification accuracy of the Chi-square method is 93.14%. Lu et al [6] used the default ARTMAP as the classifier for the multiclass cancer classification with the same data

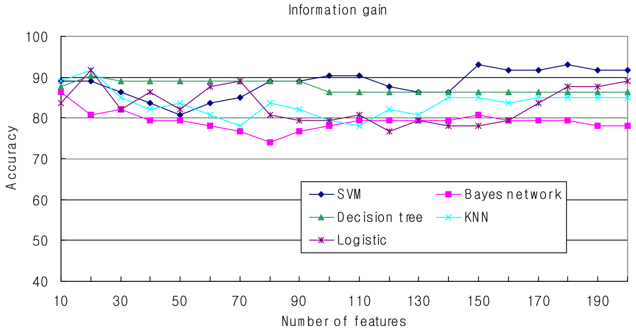

Figure 3. Classification accuracy (%) with information gain feature selection. KNN = k-nearest-neighbor; SVM, support vector machine.

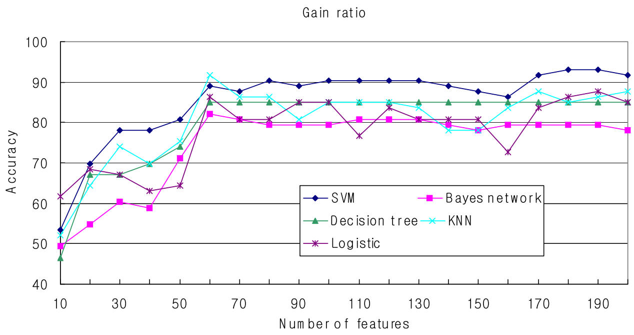

Figure 4. Classification accuracy (%) with gain ratio feature selection. KNN = k-nearest-neighbor; SVM, support vector machine.

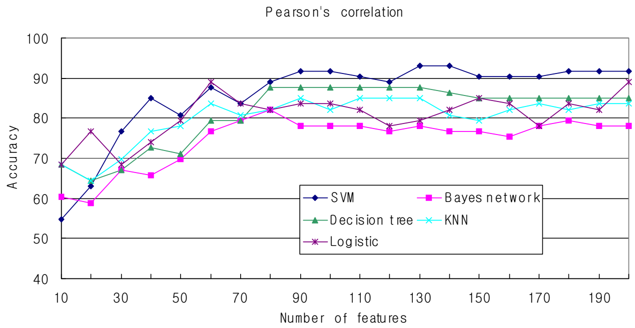

Figure 5. Classification accuracy (%) with Pearson's correlation feature selection. KNN = k-nearest-neighbor; SVM, support vector machine.

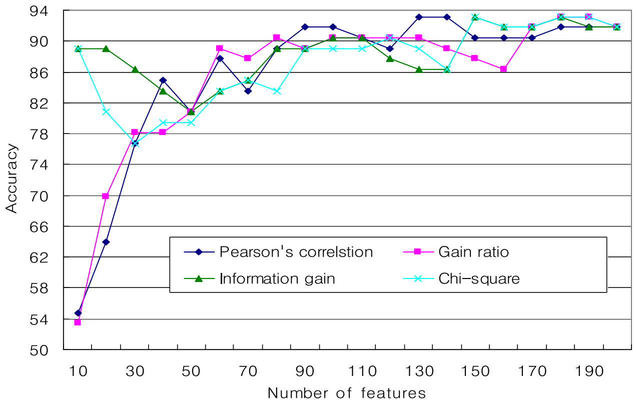

Figure 6. Classification accuracy (%) of four feature selection methods with support vector machine classifier.

Table 3. Classification accuracy (%) for support vector machine classifier.

Relationship[a]	Information gain	Chi-square
1:1 or n:1	89.04	89.04
1:n	52.05	50.68
m:n	94.52	93.14

[a] 1:1, n:1, 1:n and m:n indicate the relationship between feature and class: 1:1 and n:1 mean the high-ranking features; 1:n means the low-ranking features; m:n means both the high and low-ranking features.

set as in our work. However, the best result only has an accuracy of 88.89%. Also, as our previous work shows, in Figure 6, there is no accuracy >94%. Thus, it can be seen that feature selection with the m:n features achieved the highest classification accuracy. The results proved that it is reasonable to take the low-ranking features into consideration during cancer classification.

4. Discussion

In our work, we considered all cases (1:1, n:1, 1:n, and m:n) in cancer classification. To achieve this goal, we

Table 4. Ten high-ranking miRNAs selected by the information gain method.

Probe ID	Target sequence	MiRNA name
EAM250	AUGACCUAUGAAUUGACAGAC	hsa-miR-215
EAM330	UGUAAACAUCCUCGACUGGAAGC	hsa-miR-30a-5p
EAM105	UCCCUGAGACCCUAACUUGUGA	hsa-miR-125b
EAM348	CAUCAAAGUGGAGGCCCUCUCU	mmu-miR-291-5p
EAM190	UACCCUGUAGAACCGAAUUUGU	hsa-miR-10b
EAM288	CCCUGUAGAACCGAAUUUGUGU	mmu-miR-10b
EAM366	UUCAGCUCCUAUAUGAUGCCUUU	mmu-miR-337
EAM261	AUCACAUUGCCAGGGAUUACCAC	hsa-miR-23b
EAM260	AUCACAUUGCCAGGGAUUUCC	hsa-miR-23a
EAM381	UCGAGGAGCUCACAGUCUAGUA	rno-miR-151*

Table 5. Ten high-ranking miRNAs selected by the Chi-square method.

Probe ID	Target sequence	cMiRNA name
EAM250	AUGACCUAUGAAUUGACAGAC	hsa-miR-215
EAM190	UACCCUGUAGAACCGAAUUUGU	hsa-miR-10b
EAM288	CCCUGUAGAACCGAAUUUGUGU	mmu-miR-10b
EAM105	UCCCUGAGACCCUAACUUGUGA	hsa-miR-125b
EAM366	UUCAGCUCCUAUAUGAUGCCUUU	mmu-miR-337
EAM381	UCGAGGAGCUCACAGUCUAGUA	rno-miR-151*
EAM303	UACAGUAGUCUGCACAUUGGUU	hsa-miR-199a*
EAM336	AGGCAGUGUAGUUAGCUGAUUGC	hsa-miR-34c
EAM339	CACCCGUAGAACCGACCUUGCG	hsa-miR-99b
EAM260	AUCACAUUGCCAGGGAUUUCC	hsa-miR-23a

selected the high- and low-ranking features using information gain and Chi-square feature selection, respectively. Our work has proved the usefulness of the m:n features in cancer classification because the results showed that considering both the high- and low-ranking miRNAs can lead to higher classification accuracy than just considering the high-ranking miRNAs. Furthermore, the selected low-ranking miRNAs in Tables 6 and 7 provide cancer researchers with some very useful information for further research analysis regarding their function in human cancer. However, our work has one shortcoming; although multiple experiments have been done to find a relatively good number of the m:n features for analysis, it is difficult to determine the best number of selected features.

In future work, we will do our best to discover some feature selection algorithms that can elect the

Table 6. Seventeen low-ranking miRNAs selected by the information gain method.

Probe ID	Target sequence	MiRNA name
EAM247	UAACAGUCUCCAGUCACGGCC	hsa-miR-212
EAM252	UACUGCAUCAGGAACUGAUUGGAU	hsa-miR-217
EAM254	UGAUUGUCCAAACGCAAUUCU	hsa-miR-219
EAM259	UGUCAGUUUGUCAAAUACCCC	hsa-miR-223
EAM283	UUCCCUUUGUCAUCCUUUGCCU	mmu-miR-211
EAM293	CAUCCCUUGCAUGGUGGAGGGU	hsa-miR-188
EAM306	UACUCAGUAAGGCAUUGUUCU	mmu-miR-201
EAM308	UGGAAUGUAAGGAAGUGUGUGG	hsa-miR-206
EAM309	GCUUCUCCUGGCUCUCCUCCCUC	mmu-miR-207
EAM328	CAGUGCAAUAGUAUUGUCAAAGC	hsa-miR-301
EAM331	UGUAAACAUCCUUGACUGGA	hsa-miR-30e
EAM337	CAAAGUGCUGUUCGUGCAGGUAG	hsa-miR-93
EAM340	CUAUACGACCUGCUGCCUUUCU	mmu-let-7d*
EAM341	CAAAGUGCUAACAGUGCAGGUA	mmu-miR-106a
EAM346	CUCAAACUAUGGGGGCACUUUUU	mmu-miR-290
EAM352	AAAGUGCUUCCCUUUUGUGUGU	mmu-miR-294
EAM361	CCUCUGGGCCCUUCCUCCAGU	hsa-miR-326

Table 7. Seventeen low-ranking miRNAs selected by the Chi-square method.

Probe ID	Target sequence	MiRNA name
EAM247	UAACAGUCUCCAGUCACGGCC	hsa-miR-212
EAM252	UACUGCAUCAGGAACUGAUUGGAU	hsa-miR-217
EAM254	UGAUUGUCCAAACGCAAUUCU	hsa-miR-219
EAM259	UGUCAGUUUGUCAAAUACCCC	hsa-miR-223
EAM283	UUCCCUUUGUCAUCCUUUGCCU	mmu-miR-211
EAM290	UGGACGGAGAACUGAUAAGGGU	hsa-miR-184
EAM293	CAUCCCUUGCAUGGUGGAGGGU	hsa-miR-188
EAM308	UGGAAUGUAAGGAAGUGUGUGG	hsa-miR-206
EAM309	GCUUCUCCUGGCUCUCCUCCCUC	mmu-miR-207
EAM324	CAUUGCACUUGUCUCGGUCUGA	hsa-miR-25
EAM328	CAGUGCAAUAGUAUUGUCAAAGC	hsa-miR-301
EAM331	UGUAAACAUCCUUGACUGGA	hsa-miR-30e
EAM337	CAAAGUGCUGUUCGUGCAGGUAG	hsa-miR-93
EAM340	CUAUACGACCUGCUGCCUUUCU	mmu-let-7d*
EAM341	CAAAGUGCUAACAGUGCAGGUA	mmu-miR-106a
EAM346	CUCAAACUAUGGGGGCACUUUUU	mmu-miR-290
EAM352	AAAGUGCUUCCCUUUUGUGUGU	mmu-miR-294

appropriate m:n features automatically. In addition, we will try to use this idea to test for other types of data in addition to the miRNA expression data.

Conflicts of interest

All contributing authors declare no conflicts of interest.

Acknowledgments

This research was supported by the research grant of Chungbuk National University in 2013, the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2013R1A2A2A01068923), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2008-0062611), and the MSIP (Ministry of Science, Information and Communication Technology (ICT) and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2014-H0301-14-1022) supervised by the NIPA (National IT Industry Promotion Agency).

References

1. Mraz M, Pospisilova S. MicroRNAs in chronic lymphocytic leukaemia. From causality to associations and back. Exp Rev Hematol 2012 Dec;5(6):579—81.
2. He L, Thomson JM, Hemann MT, et al. A microRNA polycistron as a potential human oncogene. Nature 2005 Jun;435(7043):828—33.
3. Mraz M, Pospisilova S, Malinova K, et al. MicroRNAs in chronic lymphocytic leukemia pathogenesis and disease subtypes. Leuk Lymphoma 2009 Mar;50(3):506—9.
4. Lee RC, Feinbaum RL, Ambros V. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 1993 Dec;75(5):843—54.
5. Cunningham JM, Oberg AN, Borralho PM, et al. Evaluation of a new high-dimensional miRNA profiling platform. BMC Med Genomics 2009 Aug;2(57):1—14.
6. Lu J, Getz G, Miska EA, et al. MicroRNA expression profiles classify human cancers. Nature Jun 2005;435(7043):834—8.
7. Zheng Y, Chee KK. Cancer classification with microRNA expression patterns found by an information theory approach. J Computer 2006 Aug;1(5):30—9.
8. Xu R, Xu J, Wunsch 2nd DC. MicroRNA expression profile based cancer classification using Default ARTMAP. Neural Networks 2009 Jul—Aug;22(5—6):774—80.
9. Kim KJ, Cho SB. Exploring features and classifiers to classify microRNA expression profiles of human cancer. Neural Information Processing 2012;6444:234—41.
10. Moller HG, Rasmussen AP, Andersen HH, et al. A systematic review of microRNA in glioblastoma multiforme: micro-modulators in the mesenchymal mode of migration and invasion. Mol Neurobiol 2013 Feb;47(1):131—44.
11. Hall M. Correlation-based feature selection for machine learning [PhD Thesis]. New Zealand: Department of Computer Science, Waikato University; 1999.
12. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent System and Technology 2011 Apr;(2):1—27.
13. Xu JH, Li F, Sun QF. Identification of microRNA precursors with support vector machine and string kernel. Genomics Proteomics Bioinformatics 2008 Jun;6(2):121—8.
14. Campos LM, Cano A, Castellano JG, et al. Bayesian networks classifiers for gene-expression data. 11th International Conference on Intelligent Systems Design and Applications (ISDA); 2011 Nov. p. 1200—6.
15. Mishra AK, Chandrasekharan H. Analysis and classification of plant microRNAs using decision tree based approach. In: ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India, vol. II; 2014. p. 105—10.
16. Jordan MI. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. Advances in neural information processing systems 2002;14:841—8.