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	