Computers in Biology and Medicine 80 (2017) 39–44

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Multiclass cancer classification using a feature subset-based ensemble from microRNA expression profiles

Yongjun Piao[a], Minghao Piao[b], Keun Ho Ryu[a,*]

[a] Database/Bioinformatics Laboratory, College of Electrical & Computer Engineering, Chungbuk National University, Cheongju, 28644, South Korea
[b] Department of Computer Engineering, Dongguk University Gyoengju Campus, 38066, South Korea

ARTICLE INFO

Keywords:
Cancer classification
Data mining
Ensemble learning
miRNA expression

ABSTRACT

Cancer classification has been a crucial topic of research in cancer treatment. In the last decade, messenger RNA (mRNA) expression profiles have been widely used to classify different types of cancers. With the discovery of a new class of small non-coding RNAs; known as microRNAs (miRNAs), various studies have shown that the expression patterns of miRNA can also accurately classify human cancers. Therefore, there is a great demand for the development of machine learning approaches to accurately classify various types of cancers using miRNA expression data. In this article, we propose a feature subset-based ensemble method in which each model is learned from a different projection of the original feature space to classify multiple cancers. In our method, the feature relevance and redundancy are considered to generate multiple feature subsets, the base classifiers are learned from each independent miRNA subset, and the average posterior probability is used to combine the base classifiers. To test the performance of our method, we used bead-based and sequence-based miRNA expression datasets and conducted 10-fold and leave-one-out cross validations. The experimental results show that the proposed method yields good results and has higher prediction accuracy than popular ensemble methods. The Java program and source code of the proposed method and the datasets in the experiments are freely available at https://sourceforge.net/projects/mirna-ensemble/.

1. Introduction

Cancer is a class of complex genetic diseases that are characterized by out-of-control cell growth. Cancer classification has been a crucial topic of research in cancer treatment. In the last decade, mRNA expression data have been widely used to classify different types of human cancers [1]. Various machine learning approaches have been developed [2–5] to reduce the dimensionality of mRNA expression data and improve the classification accuracy.

With the discovery of a class of small non-coding RNAs, known as microRNAs (miRNAs), the expression patterns of these molecules have attracted the attention of many researchers. miRNAs play important regulatory roles in biological processes such as development, cell proliferation, differentiation and apoptosis [6,7] by pairing the mRNA of protein-coding genes with the transcriptional or post-transcriptional regulation of their expression [8]. miRNAs have emerged as highly tissue-specific biomarkers that function as tumor suppressors and oncogenes. Furthermore, several studies have shown that the expression patterns of miRNAs are heterogeneous in different human cancers [9–11]. Therefore, there is a great demand for developing machine learning approaches to accurately classify various types of cancers from miRNA expression data. Moreover, next-generation sequencing technology is increasingly used to quantify miRNA expression levels (miRNA-seq) as an alternative to microarrays. The classification model should also be scalable to such types of 'digital' expression data.

It is well known that ensembles of classifiers can improve the prediction accuracy by constructing a set of base classifiers from the training data and performing classification by combining the results of each base classifier. Several methods to construct an ensemble have been developed [12,13], such as instance subset-based approaches (i.e. bagging and boosting) and feature subset-based approaches (i.e. random forests). However, there are some difficulties in using the instance subset-based approach for miRNA expression data classification because the main characteristic of the data is that they lack training samples compared with dimensions. The basic idea of a feature subset-based ensemble is simply to give each classifier a different projection of the training set [14]. A feature subset-based ensemble has several advantages: i) automatically removes irrelevant and redundant features, ii) it performs fast because of the reduced size of the input

* Corresponding author.
E-mail addresses: pyz@dblab.chungbuk.ac.kr (Y. Piao), myunghopark@gmail.com (M. Piao), khryu@dblab.chungbuk.ac.kr (K.H. Ryu).

http://dx.doi.org/10.1016/j.compbiomed.2016.11.008
Received 9 September 2016; Received in revised form 15 November 2016; Accepted 20 November 2016

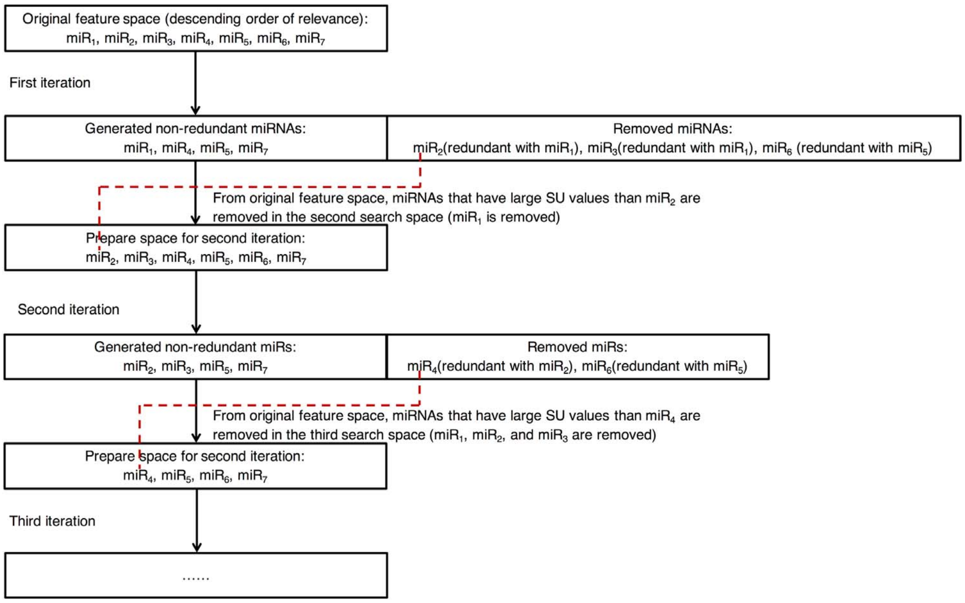
Original feature space (descending order of relevance):
miR$_1$, miR$_2$, miR$_3$, miR$_4$, miR$_5$, miR$_6$, miR$_7$
First iteration
Generated non-redundant miRNAs:
miR$_1$, miR$_4$, miR$_5$, miR$_7$
Removed miRNAs:
miR$_2$(redundant with miR$_1$), miR$_3$(redundant with miR$_1$), miR$_6$ (redundant with miR$_5$)
From original feature space, miRNAs that have large SU values than miR$_2$ are removed in the second search space (miR$_1$ is removed)
Prepare space for second iteration:
miR$_2$, miR$_3$, miR$_4$, miR$_5$, miR$_6$, miR$_7$
Second iteration
Generated non-redundant miRs:
miR$_2$, miR$_3$, miR$_5$, miR$_7$
Removed miRs:
miR$_4$(redundant with miR$_2$), miR$_6$(redundant with miR$_5$)
From original feature space, miRNAs that have large SU values than miR$_4$ are removed in the third search space (miR$_1$, miR$_2$, and miR$_3$ are removed)
Prepare space for second iteration:
miR$_4$, miR$_5$, miR$_6$, miR$_7$
Third iteration
……

Fig. 1. Example of the generation of multiple miRNA subsets.

space, and iii) it reduces the correlations among the classifiers. Generating feature subsets for an ensemble can be viewed as multiple runs of feature selection procedures. In the past, several feature selection techniques have been proposed, such as information gain, gain ratio, and correlation coefficients. However, these methods are designed to identify a best single-feature subset, which is not suitable for direct application to ensemble generation. Moreover, most of them do not consider the interactions among the features.

In this article, we propose a feature subset-based ensemble method to classify multiple tissues with miRNA expression data. First, we suggest an miRNA subset generation method based on the relevance of miRNAs to cancers and the interactions among miRNAs. Then, a number of classifiers learn from the generated subsets. Finally, the results of each classifier are combined with the average probability of the classifiers. It is observed that the proposed ensemble method obtains promising classification accuracy compared with other ensemble methods.

2. Method

An ensemble method with multiple independent feature subsets is proposed to classify various cancer types. Note that from the data mining viewpoint, an miRNA can be considered as a feature for classification purposes. The proposed method has three major steps: i) generation of multiple miRNA subsets based on the correlations among the miRNAs; ii) learning of the model from each miRNA subset using a machine learning algorithm as a base classifier; iii) combination of the results of each classifier by averaging the probabilities. In the next sections, we will illustrate each step in detail.

2.1. Generation of multiple miRNA subsets

There is no doubt that the major factor to achieve better ensemble performance is the diversity of each ensemble member. The basic idea of our subset generation method is obtaining the diversity by identifying relevant features and putting redundant features into different base classifiers. Various studies have demonstrated that the symmetrical uncertainty (SU) is a good measure to identify both relevant and redundant features [15–17]. The SU is a correlation measure of a random variable based on the information-theoretical concept of entropy. The SU between any pair of features or a feature and class can be calculated as follows:

$$\mathrm{IG}(X|Y) = H(X) - H(X|Y) \tag{1}$$

$$\mathrm{SU}(X, Y) = 2{*}\mathrm{IG}(X|Y)/(H(X) + H(Y)) \tag{2}$$

where IG(X|Y) is the information gain of X after observing variable Y, and H(X) and H(Y) are the entropy values of variables X and Y, respectively. The SU value is 0–1, where 1 indicates complete correlation and 0 indicates no correlation. To the best of our knowledge, FCBF [18] was the first method to define feature relevance and redundancy using SU.

Definition 1. (Relevant Feature) A feature X is relevant if the SU value to the class, which is denoted as SU(X,C), is larger than a user-defined threshold.

Definition 2. (Redundant Feature) Relevant features X and Y are redundant if their SU, which is denoted as SU(X,Y), is larger than min(SU(X,C), SU(Y,C)).

Given a discretized miRNA expression dataset D with m samples and n miRNAs (miR$_1$, miR$_2$..., miR$_n$), the proposed method first searches all relevant miRNAs and sorts them in descending order of relevance. Note that the irrelevant miRNAs will no longer be considered. Then an miRNA miR$_t$ is selected as the starting point to generate a subset. Similar to FCBF, the proposed method finds all redundant miRNAs and forms a non-redundant feature subset by removing the redundant miRNAs with miR$_t$. The next key point is how to form the input space to generate the next non-redundant subset. Between two redundant miRNAs, the less relevant miRNA may also produce a competitive result when the combinations of the miRNAs are considered. Therefore, we use the less relevant miRNA as a starting point for the next search. For ease of understanding, let us consider a subset search procedure with 7 miRNAs, as illustrated in Fig. 1. Suppose that all miRNAs are already sorted in descending order of their relevance; then, the most relevant miRNA, miR$_1$, is selected as a starting point to generate the first subset. Assume that miR$_2$ and miR$_3$ are redundant with miR$_1$, and miR$_6$ is redundant with miR$_5$; then, miR$_2$, miR$_3$, and miR$_6$ are removed. The remaining subset, which includes miR$_1$, miR$_4$, miR$_5$, and miR$_7$, is the generated subset in the first search. In the second iteration, we select the most relevant miRNA among the removed miRNAs (in this case, miR$_2$) as the starting point. In addition, it is not necessary to analyze the entire feature space again because the