Gene 518 (2013) 159–163

Contents lists available at SciVerse ScienceDirect

Gene

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

Classifying subtypes of acute lymphoblastic leukemia using silhouette statistics and genetic algorithms

Tsun-Chen Lin [a,*], Ru-Sheng Liu [b], Ya-Ting Chao [c], Shu-Yuan Chen [b]

[a] Department of Computer Science and Engineering, Dahan Institute of Technology, Hualien, 970, Taiwan, ROC
[b] Department of Computer Science and Engineering, Yuan Ze University, Taoyuan, 32026, Taiwan, ROC
[c] Graduate School of Biotechnology and Bioinformatics, Yuan Ze University, Taoyuan, 32026, Taiwan, ROC

ARTICLE INFO

Available online 10 December 2012

Keywords:
Genetic algorithm
Silhouette statistics
Microarray
Classification
Pediatric acute lymphoblastic leukemia

ABSTRACT

Correct classification and prediction of tumor cells is essential for a successful diagnosis and reliable future treatment. In this study, we aimed at using genetic algorithms for feature selection and proposed silhouette statistics as a discriminant function to distinguish between six subtypes of pediatric acute lymphoblastic leukemia by using microarray with thousands of gene expressions. Our methods have shown a better classification accuracy than previously published methods and obtained a set of genes effective to discriminate subtypes of pediatric acute lymphoblastic leukemia. Furthermore, the use of silhouette statistics, offering the advantages of measuring the classification quality by a graphical display and by an average silhouette width, has also demonstrated feasibility and novelty for more difficult multiclass tumor prediction problems.

1. Introduction

The analysis of gene expression profiles that serve as molecular signatures for tumor/cancer classification has become a highly challenging research topic in bioinformatics. Generally, the classification of microarray data of cancers can be delineated into two tasks: gene selection and classification. Gene selection searches class discriminant genes for classification, from thousands of gene expression profiles. Classification requires the construction of a model, which processes input gene patterns representing objects, and predicts the class or category associated with the objects under consideration.

Due to the special characteristics of microarray data's classification problem, that is the very small samples in an extremely high dimensional input gene space, many computational algorithms, based on rank based gene selection schemes (gene vs. ideal gene) have been successfully applied to microarray data classification task, including the use of hierarchical clustering (Alizadeh et al., 2000), support vector machine (Furey et al., 2000), neighborhood/SOM analysis (Golub et al., 1999), and artificial neural networks (Khan et al., 2001). In fact, there are more types of cancers, and potentially even more subtypes, and when the heterogeneity of cancers is still the most significant problem in the practical management of the individual patient, the development of methods for multiclass classification problems has become necessary (Li et al., 2004). Instead of ranking genes for feature selection, many approaches based on genetic algorithms have been proposed to evaluate candidate genes in chromosomes to be used as input data for classifiers (Deutsch, 2003; Li et al., 2001; Lin et al., 2006; Liu et al., 2005; Ooi and Tan, 2003). These methodologies, such as the maximum likelihood method of GA/MLHD and the silhouette statistics of GASS, have shown their superiorities to capture informative genes and to improve the prediction accuracy in the multiclass microarray classification problems, especially when the number of classes is more than five.

Pediatric Acute Lymphoblastic Leukemia with many cancer subtypes is the most common type of leukemia in children. In this paper, we extend our previous work of GASS to classify the gene expression profiles of acute lymphoblastic leukemia. In the results of our experiments, this methodology has exhibited 100% classification accuracy, and needed a less number of discriminating genes compared to reported techniques based on the same dataset. Moreover, we also introduced the new application of silhouette statistics in classification quality analysis, whereas it was originally published for the quality of clustering analysis (Rousseeuw, 1987). This is a plausible improvement in the multiclass microarray classification problem and may be a useful method in cancer diagnosis.

2. Methods

2.1. Classification based on silhouette statistics

In this section, we describe the main discriminant method that we will use in this paper. For tumor pattern classification, our definition

Abbreviations: BCR–ABL, breakpoint cluster region Abelson (ABL); E2A–PBX1, E2A–PBX1 fusion gene; Hyperdiploid > 50, hyperdiploid (> 50 chromosome); MLL, myeloid/lymphoid leukemia; T-ALL, T-lineage acute lymphoblastic leukemia; TEL–AML1, TEL–AML1 chimeric fusion gene.

* Corresponding author. Tel.: +886 3 8210872; fax: +886 3 8266588.
E-mail address: lintsunc@ms01.dahan.edu.tw (T.-C. Lin).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2012.11.046

Table 1
Distance metrics.

Metrics	Formula
Euclidean	$d_E(\vec{e}_i,\vec{e}_j)=\left\{\sum_G\left(\vec{e}_{Gi}-\vec{e}_{Gj}\right)^2\right\}^{1/2}$
Minkowski	$d_{Mk}(\vec{e}_i,\vec{e}_j)=\left\{\sum_G\left(\vec{e}_{Gi}-\vec{e}_{Gj}\right)^{\lambda}\right\}^{1/\lambda},\ \lambda=\infty$
1-Pearson	$d_P(\vec{e}_i,\vec{e}_j)=1-\frac{\sum_G(\vec{e}_{Gi}-\bar{\vec{e}}_x)(\vec{e}_{Gj}-\bar{\vec{e}}_y)}{\left\{\sum_G(\vec{e}_{Gi}-\bar{\vec{e}}_x)^2\right\}^{1/2}\left\{\sum_G(\vec{e}_{Gj}-\bar{\vec{e}}_y)^2\right\}^{1/2}}$
1-Spearman	$d_{Sp}(\vec{e}_i,\vec{e}_j)=1-\frac{6\sum(d_i-d_j)^2}{G(G^2-1)}$
	d_i, d_j are the rank vectors of $\vec{e}_i$ and $\vec{e}_j$, respectively.

Note: (1) The vector $-e_i=(0,2,-2,3,-3)$ is transformed into the rank vector $-d_i=(3,4,2,5,1)$ where the smallest value has rank 1 and the largest number has rank 5. (2) $\bar{e}_i=(1/G)\sum_G e_i^{(G)}$.
Distance-based measures: Euclidean, Minkowski; correlation-based measures: 1-Spearman, 1-Pearson.

starts by assuming that we are given a dataset D. Let $D=\{(\vec{e}_j, l_j)$, for $j=1\ldots m\}$ be a set of m number of samples, where $\vec{e}_j=(e_{j1},e_{j2},\ldots,e_{jG})^t$ is the vector of tumor pattern for the ith sample that describes expression levels of G number of predictive genes, and $l_j\in L=\{c_1,c_2,\ldots,c_q\}$ is the class label associated with $\vec{e}_j$. Our discriminant function based on the silhouette statistics is then defined as

$$Sil\left(\vec{e}_i\right)=\frac{b\left(\vec{e}_i\right)-a\left(\vec{e}_i\right)}{\max\left\{a\left(\vec{e}_i\right),b\left(\vec{e}_i\right)\right\}}. \quad (1)$$

This formula begins with defining $d\left(\vec{e}_i,c_s\right)$ as the average distance of $\vec{e}_i$ to other observations of samples in the class of c_s. Then $b(\vec{e}_i)$ denotes $\min\{d\left(\vec{e}_i,c_s\right)\}$, $\vec{e}_i\in c_r$, $r\neq s$, $r\in\{1,2\ldots q\}$, q is the number of classes, and $a(\vec{e}_i)$ denotes $d\left(\vec{e}_i,c_s\right)$, $\vec{e}_i\in c_r$, $r=s$. In other words, $a(\vec{e}_i)$ is the average distance between tumor patterns of $\vec{e}_i$ and all other sample patterns in the class to which $\vec{e}_i$ belongs, and $b(\vec{e}_i)$ is the minimum average distance of $\vec{e}_i$ to objects in other classes. The $Sil(\vec{e})$ is the discriminant function, returning the discrimination score, ranging from -1 to $+1$, to indicate how well a test sample, represented by the vector of $\vec{e}$, can be assigned to its own class. Intuitively, samples with a large silhouette value are well classified, and those with a small silhouette value tend to lie between classes, and those with a negative value are poorly classified. In order to classify samples into their own classes without a negative silhouette value, we set $Sil(\vec{e})>0$ as a criterion for each sample to be correctly classified. This means that once the returning value is less than zero, we say that the corresponding sample is misclassified under the discriminant variable of $\vec{e}$. Therefore, the classification rule can be written as

$$C\left(\vec{e}_i\right)=l_i,\ \text{where } Sil\left(\vec{e}_i\right)>\theta,\ \theta=0 \quad (2)$$

where θ is the silhouette threshold value. Note that the classification rule can also be used to predict the labels of novel samples. For a novel sample, its label should be assumed to be from C_1 to C_q and the corresponding silhouette value should be calculated by Eq. (1). Since there exists only one class deserving the minimum average distance for the novel sample, only one positive silhouette value can be obtained. In contrast, in our experiments if a novel sample is assigned to the class that returns a positive silhouette value causing the predicted label to be different from the actual class label, we can state that a misclassification has occurred. Here, we may also find that the efficiency of silhouette statistics depends on two factors: (1) the distance metric used in silhouette statistics, and (2) the sample pattern $\vec{e}_j$. For the distance metric, we implement two groups of two distance metrics to compare the effects on silhouette statistics. The first one is the distance-based measures, and the other one is the correlation-based measures (Table 1). Besides, we also discuss how the pattern of $\vec{e}_j$ can be chosen by genetic algorithms in the next section.

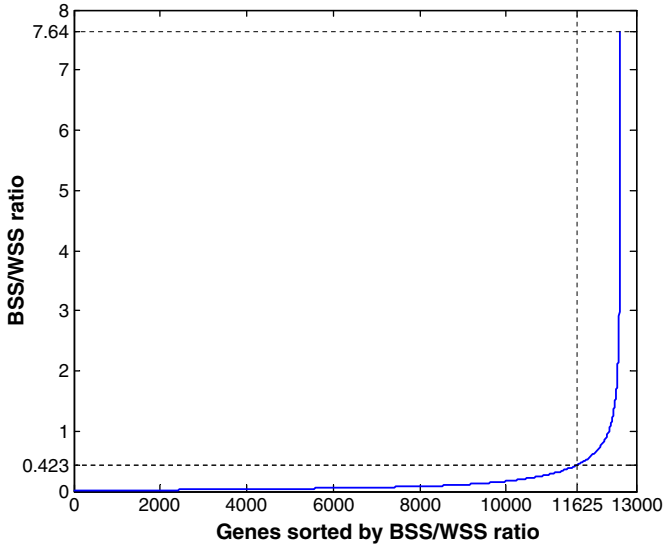

Fig. 1. This figure showed that roughly 1000 genes were more highly correlated with the class distinction than would be expected by chance.

2.2. Genetic algorithm for gene selection

In order to select an optimal subset of features from a large feature space, we employ the GA approach. The genetic algorithms use two selection methods including stochastic universal sampling (SUS) and roulette wheel selection (RWS). In addition, two tuning parameters, P_c: crossover rate and P_m: mutation rate, are used to tune one-point and uniform crossover operations to evolve the population of individuals in the mating pool. The format of chromosomes used to carry subsets of genes is defined by the string S_i, $S_i=[G,g_1,g_2,\ldots g_{Gmax}]$, where g_i is the expression level of gene i and G denotes the number of predictive genes to form sample patterns in the classification and is limited to a predefined range from G_{min} to G_{max}. In our algorithms, we will try as many chromosomes as possible to choose the optimal gene subset by scoring those chromosomes using the fitness function of $f(S_i)=(1-E_t)\times 100$, where E_t means the training error rate of leave-one-out cross validation (LOOCV) test. In order to have an unbiased estimation of initial gene pools, our algorithms will set 20 gene pools to run the following steps.

Step 1: For each gene pool, the evolution process will execute 100 generations and each generation will evolve 100 chromosomes in which the size of genes will range from $G_{min}=15$ to $G_{max}=25$.
Step 2: According to the gene indices in each chromosome, only the first G genes are picked from g_1, $g_2\ldots g_{Gmax}$ to form sample patterns for classification. In other words, the dataset is then represented by a matrix $X_{G\times m}$ form with rows for the G genes and columns for the m samples.
Step 3: In order to estimate the fitness score for each chromosome, the training dataset $X_{G\times P}$ of P training samples is used in the following program to evaluate how well those samples can be correctly classified under silhouette statistics.

1. FOR each chromosome S_i //$i=1$ to 100
2. FOR each training sample with class label l_j
3. Build up discriminant model with the remaining training samples for LOOCV test
4. IF $(Sil(-e_j)<0)$