Applied Soft Computing 48 (2016) 683–690

Contents lists available at ScienceDirect

Applied Soft Computing

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

Gene discretization based on EM clustering and adaptive sequential forward gene selection for molecular classification

Hung-Yi Lin

Department of Distribution Management, Taichung University of Science and Technology 129, Sanmin Rd., Sec. 3, Taichung, TAIWAN, R.O.C.

ARTICLE INFO

Article history:
Received 9 June 2015
Received in revised form 10 May 2016
Accepted 6 July 2016
Available online 14 July 2016

Keywords:
Gene discretization
EM clustering
Sequential forward selection
Molecular classification

ABSTRACT

The mismatch in gene dimension as opposed to sample dimension poses a great challenge for many modelling problems in bioinformatics. Feature selection in immense quantities of high-dimensional data for molecular classification renews the tasks to the modern data mining techniques. The advent of microarray datasets pushed research in bioinformatics to a new boundary in the last decade. Many bioinformatics applications necessiate feature selection or dimensionality reduction techniques for identifying informative genes or selecting subset of genes with discrimination power. Here, gene discretization based on EM clustering for complexity simplification and better discrimination capability is employed. Then, an adaptive sequential forward search algorithm for the exploration of distinct subsets of genes with discrimination power is proposed. By monitoring the information gain acquired from a collection of selected features, we are able to predict distinction between multiple subclasses without previous knowledge of these subclasses. Experimental results demonstrate the feasibility of cancer classification based solely on the discretized gene expression monitoring, completely independent of previous biological knowledge.

1. Introduction

The collection of biological data via clinical courses or diagnostic results is inefficient and unable to get abreast of the speed of information technologies. Terabytes of molecular data of gene expression levels are nowadays effortless to be generated and continuously accumulated for many purposes. Many study issues including biomarker discover, prognostic test, pathology, disease subclass discovery/characterization, class prediction for new instance, instance grouping/clustering, critical gene authentication, pharmacological test, toxicogenomics, and so on, are interested to the biological or biomedical fields. In addition to the characteristics of volume, value and velocity, the variety of gene expression profiles derived from the biology of natures deeply fascinates many research attention. In order to utilize the implicitly useful information while avoiding the winding of noisy information, the raw data in microarrays necessitate preprocessing before carrying on further handles. Data simplification and data reduction are two major aims when preprocessing tremendous amount of data. Data simplification methods must reduce data complexity and retain the inherent informative message. Simplification handles do not change data scale while data reduction methods have to remove the useless message and in turn decrease data scale. Accurate data simplification collaborated with precise data reduction can lead to the high quality of data analyses. In this paper, gene discretization and gene subset searching are respectively dealt with EM clustering and the adaptive sequential forward selection method for the achievements of data simplification and data reduction.

The unified discrimination effect corresponding to a subset of features depends on the coregulation and cofunction between the selected features. To authenticate the relevance of the selected features to the target class variable, many feature subset selection methods were developed. *Filter* and *wrapper* are two broad categories of optimal feature subset selection. In filter approaches, features are scored and ranked based on statistical criteria. Frequently used filter methods include *t*-test [16], Chi-Squared (χ^2) test [18,23], Pearson correlation coefficients [4], correlation-based feature selection (CFS) [13], mutual information [22,26] and principal component analysis [21]. Feature selection filters using low-complexity relevance and redundancy criteria are proposed in [10].

In wrapper approaches [19], feature selection is "wrapped" in a certain learning algorithm. Based on feature search strategy, wrapper methods can broadly be classified into *greedy* and *stochastic*. Sequential backward selection (SBS) and sequential forward selection (SFS) [7] are two most commonly used wrapper methods using greedy search strategy. SBS is robust to interaction problems but

E-mail addresses: linhy@nutc.edu.tw, linhy55@gmail.com

http://dx.doi.org/10.1016/j.asoc.2016.07.015
1568-4946/

sensitive to multicollinearity. On the other hand, SFS is robust to multicollinearity problems but sensitive to feature interaction. SFS and SBS both rely on the monotonicity assumption and cause the problem of single-track search. Stochastic algorithms such as ant colony optimization (ACO), genetic algorithm (GA), particle swarm optimization (PSO) are developed for solving large scale combinatorial problems. However, these algorithms are still computationally expensive.

Many studies [14,15,24,29] propose hybrid approaches taking advantages of filter and wrapper methods. The idea behind the hybrid methods is that filter methods first reduce feature space and then wrapper methods are applied to find the optimal subset of features from the reduced feature space. Greedy or exploratory searching within an effective number of features makes feature selection faster and more economic. However, feature interactions may persecute hybrid methods in terms of classification accuracy because a relevant feature in isolation may appear no more discriminating than an irrelevant one in the presence of feature interactions. The novel hybrid method proposed in this paper takes feature interactions into account. To ensure the interactions of the selected features can ameliorate the discriminative effect of single features, our method adopts the criterion of information gain in measuring the relevance of subset features to the target variable.

Our contributions in this paper are twofold. First, although cluster analyses have been widely used for many applications, discretizing feature values with EM clustering algorithm and maximizing the inter-dependency between features and the target variable is first presented. Authenticating various subsets of informative features using an adaptive sequential forward selection method is the second aspect. Based on our designs, discrimination power for the classification task is significantly boosted and different gene subsets are produced for the consistency of classification task.

This paper is organized as follows. Section 2 reviews the past related works of feature discretization and EM clustering algorithm. Section 3 presents the effectiveness of discrimination power boosting and proposes our adaptive sequential forward gene selection algorithm. The experimental and analytical results are presented in Section 4. The further discussions including discrimination analyses, over-fitting problem, and complexity analyses are presented in Section 5. Finally, concluding remarks are given in the last section.

2. Related works and studies

The measure of gene expression levels depends on many factors. Different uncertainties arise from different situations. For example, devices calibrating problem, probing environment (temperature, humidity, brightness), hybridization of samples or chips, and so forth. Straight use of expression profile in analytical procedure takes too much impurity or controversial information into account and tends to mislead or confuse the learning model. Extracting the information of distinguishability from the raw data is more helpful and significant than directly utilizing them. Data preprocessing is the first procedure and the crucial step at the beginning when proceeding a learning model. Adequate preprocessing can intensify the valuable information while improper preprocessing can leave the pivotal elements. In this paper, data simplification with the preservation of discrimination power is regarded as the first milestone to attain concrete learning results.

As far as the learning task for molecular classification is concerned, the kinds of decisive classes are farther simple than the variety of expression profiles. Using of raw data without data simplification in classification tasks tends to result in over-computing and over-fitting problems. Over-computing problems rapidly increase training costs and cause over-fitting problems. Over-fitting problems make decision rules hard to understand and cannot generate concrete results. Feature discretization (FD) can transfer continuous genes into discrete counterparts. FD techniques [3,20] aim at finding a representation of each feature that preserves enough information for the learning task, while ignoring minor fluctuations that may be irrelevant for that task. The quality of classification accuracy and model training time are deeply influenced by the selection of FD methods. Equal-interval-binning (EIB) and equal-frequency-binning (EFB) are traditional supervised discretization methods. Highly sensitive to outliers and unknown underlying range of values are two drawbacks of EIB. Although non-uniform quantization method EFB is less sensitive to outliers, the relatively smaller number of instances causes the representativeness of counted frequency is worried. The maximum entropy algorithm [6], class-attribute interdependency algorithm [5], and statistics-based [28] obtain possible solution through supervising and optimizing a cost function. Unfortunately, many decisive factors cannot be explicitly outlined in advance of selecting the supervised discretization method. For example, binning methods necessiate explicit population parameters. Data fitness to the designated discretization criterion is unpredictable. Unknown data distribution and insufficient number of representative instances are also problematic.

In principle, discretization algorithms should maximize the interdependency between discrete attribute values and class labels while minimizing the information loss due to the discretizing procedures [30]. The similarity, proxity, and connectivity among continuous values are more meaningful than their measured values. Unsupervised techniques [8,9,27] are completely independent of previous knowledge and tend to get rid of the mentioned problems when discretizating features. In order to ensure the discrimination powers of features are retained, FD using unsupervised cluster analysis is applied in this paper.

Cluster analysis is the task of grouping a set of objects in such a way that objects in the same cluster are more similar to each other than to those in other clusters. Clustering gene expression data can be categorized into three groups [25]: 1) gene-based, 2) sample-based and 3) subspace clustering as both genes and samples are required. Gene-based clustering treats genes as objects while treats samples as features [2,17]. The intention of gene-based clustering is to group genes which imply coregulation and cofunction. High level of background noisy data vastly challenge gene-based clustering algorithms. The enormous amount of genes is also problematic because neither "great number of resulting clusters" nor "huge gene amount in a cluster" is preferred. Sample-based clustering regards samples as objects and genes as features. To find the substructure of samples, it generally necessiates only a small subset of informative genes. The mismatch in gene dimension as opposed to sample dimension becomes a bottleneck as the excessive amount of genes is involved to distiguish a quite limited number of class distinction. The resulting clusters tend to have the weak class validity or become over-partitioning. Subspace clustering [1] techniques treat genes and samples symmetrically in gene expression matrices. A single gene may be involved in multiple pathways that it may or may not be coactive under all conditions. Subspace clustering defines a sub-matrix by a subset of genes on a subset of samples for the description of different explored scenarios. Completely different from the three mentioned strategies, the unsupervised expectation-maximization clustering algorithm is employed in this paper for single gene and subset of genes discretization. Namely, instead of proceeding gene and/or sample selection, cluster analysis is purely applied for FD.

The expectation-maximization algorithm (EM) can iterate and then converge the statistical model with unobserved latent variables for finding the maximum likelihood of the observed samples. EM is particularly suitable in cases where models or systems