Expert Systems with Applications 38 (2011) 12151–12159

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Multiple-kernel SVM based multiple-task oriented data mining system for gene expression data analysis ☆

Zhenyu Chen [a,*], Jianping Li [b], Liwei Wei [b], Weixuan Xu [b], Yong Shi [c,d]

[a] Department of Management Science and Engineering, School of Business Administration, Northeastern University, Shenyang 110819, China
[b] Institute of Policy & Management, Chinese Academy of Sciences, Beijing 100080, China
[c] Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing 100080, China
[d] College of Information Sciences and Technology, University of Nebraska at Omaha, Omaha, NE 68118, USA

ARTICLE INFO

Keywords:
Support vector machine
Multiple-kernel learning
Feature selection
Data fusion
Decision rule
Associated rule
Subclass discovery
Gene expression

ABSTRACT

Gene expression profiling using DNA microarray technique has been shown as a promising tool to improve the diagnosis and treatment of cancer. Recently, many computational methods have been used to discover maker genes, make class prediction and class discovery based on gene expression data of cancer tissue. However, those techniques fall short on some critical areas. These included (a) interpretation of the solution and extracted knowledge. (b) Integrating various sources data and incorporating the prior knowledge into the system. (c) Giving a global understanding of biological complex systems by a complete knowledge discovery framework. This paper proposes a multiple-kernel SVM based data mining system. Multiple tasks, including feature selection, data fusion, class prediction, decision rule extraction, associated rule extraction and subclass discovery, are incorporated in an integrated framework. ALL-AML Leukemia dataset is used to demonstrate the performance of this system.

1. Introduction

DNA microarray technology makes it possible to measure the simultaneous expression levels of thousands of genes in a single experiment and allow us to investigate the biological molecular state of a living cell. There are numerous potential applications for DNA microarray technology. Especially, it provides valuable insights into the molecular characteristics of cancers. Utilizing the gene expression data might be the most direct way to improve the diagnosis accuracy and discover the therapeutic drug target (Alizadeh, Eisen, & Davis, 2000; Alon et al., 1999; Golub, Slonim, & Tamayo, 1999; Matthias, Anthony, & Nikola, 2003; Yeoh, Ross, & Shurtleff, 2002).

Many machine learning and statistical methods have been used to discover the potential knowledge from gene expression data of cancer tissue and try to give new insights into the diagnosis and treatment of cancers. Most of the literatures for tissue gene expression data analysis focus on the following three issues: Gene identification, class discovery, and class prediction.

Gene identification is to select genes with powerful explanation capacity to a specific task such as classification. Gene identification can improve the accuracy of classifiers and reduce the computational costs. More importantly, a small subset of biological relevant genes may be useful for understanding the underlying mechanism of cancer and designing less expensive experiments (Niijima & Kuhara, 2006). Usually, all the methods proposed to tackle with gene identification yield two basic categories: Filter and wrapper methods (Inza, Larranaga, Blanco, & Cerrolaza, 2004). In general, the wrapper methods can obtain better predictive accuracy than the filter methods, but they require more computational time. Besides, gene identification algorithms may be categorized in another way: Gene ranking and gene subset selection. Gene ranking is the most commonly used technique. It evaluates each gene individually with respect to a certain criterion and assigns a score reflecting the ability to distinguish between various classes of samples. The measures based on machine learning and statistical learning theory (Fuhrman et al., 2000; Golub et al., 1999; Su, Murali, Pavlovic, Schaffer, & Kasif, 2003), such as information gain, twoing rule, sum minority, max minority, Gini index, sum of variances, one-dimensional SVM and *t*-statistics, are the popular criteria to rank genes. It is the drawback of gene ranking methods that they evaluate each gene in isolation and ignore the gene to gene correlation. From a biological perspective, we know that groups of genes, as the atomic units, work together as pathway components and reflect the state of cell (Piatetsky-Shapiro & Tamayo, 2003). Gene subset selection, on the contrary, conducts the search for a good subset by using the classifier itself as a part of evaluation function. Considering the high computational cost in the searching process, heuristic

☆ This research has been partially supported by Grants from National Natural Science Foundation of China (#70621001,#70531040).

* Corresponding author.
E-mail addresses: zychen@mail.neu.edu.cn (Z. Chen), ljp@casipm.ac.cn (J. Li), lwwei@casipm.ac.cn (L. Wei), wxu@casipm.ac.cn (W. Xu), yshi@gucas.ac.cn (Y. Shi).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.03.025

algorithms including SVM-RFE (Duan, Rajapakse, & Wang, 2005; Guyon, Weston, & Barnhill, 2002; Tang, Zhang, & Huang, 2007) and saliency analysis (Cao, Seng, Gu, & Lee, 2003) are widely used.

Class discovery usually refers to identifying previously unknown subclasses adopting the unsupervised techniques. Hierarchical clustering is a popular unsupervised tool due to its intuitive appeal and visualization properties (Eisen, Spellman, Brown, & Botstein, 1998). Another family of clustering method is greedy search-based iterative descent clustering, such as self-organizing map (SOM) (Tamayo et al., 1999), K-means clustering (Li, Weinberg, Darden, & Pedersen, 2001) and Bayesian clustering (Roth & Lange, 2004). Clustering techniques usually face some difficulties including how to choose the right number of clusters and how to evaluate the clustering results (Monti, Tamayo, & Mesirov, 2003).

Class prediction refers to the assignment of particular samples to already-defined classes which could reflect current states or future outcomes (Golub et al., 1999). The widely used methods for class prediction include decision tree (Camp & Slattery, 2002), artificial neural network (ANN) (Ando, Suguro, & Kobayashi, 2003; Tan & Pan, 2005; Tung & Quek, 2005), support vector machine (SVM) (Mao, Zhou, & Pi, 2005; Statnikov, Aliferis, & Tsamardinos, 2005) and so on.

It is the limitation of above mentioned methods that each of them can only deal with one task: Gene identification, class discovery or class prediction. However, microarray data mining needs dealing with various tasks. It is necessary to incorporate the tasks into an enlarged system (Matthias et al., 2003):

(1) Feature selection can strongly influence the performance of the methods. The most outstanding character of gene expression data is that it contains a large number of gene expression values (several thousands to tens of thousands) and a relatively small number of samples (a few dozen). It brings great challenges for the commonly used knowledge discovery methods. When the number of features far exceed the number of training samples available, most classification and clustering methods, such as decision tree, ANN and K-means, are sensitive to noise and susceptible to overfitting (Guyon et al., 2002; Monti et al., 2003;Radivojac, Chawla, & Dunker, 2004; Tung & Quek, 2005). Therefore, features selection is a necessary prior stage of the gene expression data mining (Ando et al., 2003; Matthias et al., 2003; Tan & Pan, 2005; Tung & Quek, 2005).
Besides, many features selection methods require a prefiltering stage to refine the raw feature set to achieve a satisfactory performance (Tung & Quek, 2005; Zhu, Ong, & Dash, 2007). Especially for the wrapper-based features selection methods, it is popular to propose a hybrid wrapper and filter framework to improve the performance.
(2) With the continuous emergence of new DNA data and new array technologies, how to integrate various sources data and incorporate the prior knowledge is another challenging problem (Fellenberg, 2003).
Different microarray techniques use different mechanism to measure gene expression levels. It makes the gene expression levels reported by different techniques not comparable with each other. Consequently, a challenge is integrating databases to connect this disparate information as well as performing studies to collectively analyze those datasets from diverse sources that have heterogeneous formats. It is a natural way to adapt normalization to make the gene expression values from different data sources conformable each other (Goh & Kasabov, 2003). Another way is to use statistic methods to combine the experimental results from single source data (Filkov & Skiena, 2003; Hwang et al., 2005).
(3) Considering the high cost and long time of experimental research, various tasks in the medical field including the diagnosis on a disease, the outcome prediction and drug discovery depend on the computational methods. It is necessary to design multiple-task oriented knowledge discovery system as a complete scheme.

Currently, the hybridized pipeline is a commonly used way to integrate multiple-tasks by incorporating a sequence of methods (Kim, Zhou, Morse, & Park, 2005; Matthias et al., 2003; Radivojac et al., 2004; Tan & Pan, 2005; Tung & Quek, 2005; Sethi & Leangsuksun, 2006). For each method used in one system, it is independent to set the initialized values, carry out the optimization algorithm and tune the free parameters. Although each method performs well, their integration usually does not. A small computational error may be transmitted to the following step and enlarged. It increases the uncertainty of the data mining system. For example, a filter method is used to identify the relevant genes, SVM is used to make classification and decision tree is used to extract the comprehensible rules. The selected gene subset, that is usually not optimal by the filter method, may take great effect on the performance of SVM and decision tree. Besides, good classification performance of SVM can not be take advantage of by the next step: Rule extraction by decision tree.

In this paper, a multiple-kernel SVM (MK-SVM) is proposed for multi-task oriented microarray data mining. Unlike standard SVM that is usually viewed as a "black-box", multiple-kernel SVM based gene expression data mining system is applicable to feature selection, data fusion, class prediction, decision rule extraction, associated rule extraction and subclass discovery.

This paper is organized as follows: In Section 2, we give a series of algorithm. In Section 3, we develop a methodology for multi-task oriented gene expression data mining. Section 4 presents the case study on ALL-AML leukemia dataset. Section 5 summarizes the results and draws a general conclusion.

2. Multi-kernel based SVM: Proposed algorithms

2.1. SVM: A brief introduction

We only give a brief introduction of SVM for a typical binary classification problem. The basic SVM concepts can be found in (Chen, Li, & Wei, 2007; Cristianini & Shawe-Taylor, 2000; Vapnik, 1995).

Given a set of data points $G = \{(x_i, y_i)\}_{i=1}^n, x_i \in R^m$ and $y_i \in \{+1, -1\}$. The decision function of SVM is

$$f(x) = \langle w, \phi(x) \rangle + b, \tag{1}$$

where $\phi(x)$ is a mapping of sample x from the input space to a high-dimensional feature space. $\langle \cdot, \cdot \rangle$ denotes the dot product in the feature space. The optimal values of w and b can be obtained by solving the following regularized optimization problem:

$$\min\ J(w, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i, \tag{2}$$

$$\text{s.t.} \begin{cases} y_i(\langle w, \phi(x_i) \rangle + b) \geqslant 1 - \xi_i, & i = 1, \ldots, n, \\ \xi_i \geqslant 0, \end{cases} \tag{3}$$

where ξ_i is the ith slack variable and C is the regularization parameter.

This problem is computationally solved using the solution of its dual form:

$$\max \left\{ \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j k(x_i, x_j) \right\}, \tag{4}$$