Knowledge-Based Systems 000 (2018) 1–14

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Subtype dependent biomarker identification and tumor classification from gene expression profiles

Aiguo Wang[a], Ning An[a,*], Guilin Chen[b], Li Liu[c], Gil Alterovitz[d,e]

[a] School of Computer and Information, Hefei University of Technology, Hefei, China
[b] School of Computer and Information Engineering, Chuzhou University, Chuzhou, China
[c] School of Software Engineering, Chongqing University, Chongqing, China
[d] Center for Biomedical Informatics, Harvard Medical School, Boston, USA
[e] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, USA

ARTICLE INFO

Article history:
Received 2 October 2017
Revised 19 January 2018
Accepted 26 January 2018
Available online xxx

Keywords:
Biomarker identification
Tumor subtype
Gene selection
Microarray data
Subtype dependent

ABSTRACT

Gene expression profiles are being used to categorize disease specific genes and classify different tumor subtypes at the molecular level. Due to the inherent nature of these data having high dimensionality and small sample sizes, current conventional machine learning and statistical techniques have drawbacks in achieving satisfactory predictive classification performance in clinical samples. The typical approach to handling this situation is to eliminate noisy and redundant genes from the original gene space. There are currently multiple gene selection methods available, but most of them seek to find a common subset of genes for all tumor subtypes and fail to reflect the unique characteristics of each subtype. Consequently, in this study, we propose a general framework that aims to identify subset of genes for each tumor subtype, and also give another gene selection framework that combines the obtained subtype specific gene subsets into a single gene subset. We then present a corresponding classification model for distinguishing different tumor subtypes, and implement three specific gene selection algorithms within the two frameworks. Finally, extensive experimental results on the six benchmark microarray data validate the proposed tumor subtype dependent selection process to predict and rank specific molecular biomarkers to define tumor subtypes. This new process contributes significantly to the enhancement of tumor-predictive classification performance.

1. Introduction

Tumor metastasis and subsequent mortality place a heavy social and fiscal burden on our society. Early diagnosis of tumor is more cost effective and plays a significant role in better management, treatment, and outcomes [1]. Traditional diagnostic methods include cell based observational and biochemical examination in an organ based context, both of which rely on vast and varied domain knowledge of pathological research. Guidelines and standards of care have progressed yet maintain inherent disadvantages of bias, time, and limited accuracy. Gene mutation and subsequent loss of function or alteration in molecular pathways is a defining occurrence for most metastatic events, and measuring the differential gene expression patterns in tumor cells compared with those of a normal population is increasingly accepted to diagnose cancer, define treatments, and predict outcomes in personalized cancer care plans [2].

The rapid development and wide use of microarray technology enables simultaneous measures of expression perturbations of thousands of genes under multiple experimental conditions. These early multivariate analyses have increased our capacity to identify disease genes, drug targets, and tumor subtypes [2–5]. Accordingly, various methods of analysis, including machine learning algorithms, have been created to compare gene expression profiles. The intrinsic nature of these microarray data collections is usually characterized by high dimensionality (with thousands of gene observations over time and context) and often using a small sample size of specimens or patients to limit the statistical power for clinical use [6]. This multiplicity of classifiers and data dimensionality often causes pattern profiles to be overfit and thus predictions will suffer from poor generalization capacity [7]. There are studies suggesting that there are a few important genes that are associated with a specific classification of cancer subtypes and may be (ideally) submitted for Food and Drug Administration (FDA) validation and used for diagnosis [8]. Also, the affected gene space often

* Corresponding author.
E-mail addresses: wangaiguo2546@163.com (A. Wang), ning.g.an@acm.org (N. An), glchen@chzu.edu.cn (G. Chen), dcsliuli@cqu.edu.cn (L. Liu), gil_alterovitz@hms.harvard.edu (G. Alterovitz).

https://doi.org/10.1016/j.knosys.2018.01.025
0950-7051/

consists of a large number of noisy and redundant genes, which can diminish the performance of a classifier [9–11]. For example, *k*-nearest-neighbor algorithm is sensitive to irrelevant features in classification [12]. One feasible way to mitigate this problem is to select a subset of discriminant genes from the original gene space by filtering noisy and redundant genes using effective feature selection methods [13,14].

Gene selection, also known as feature selection and variable selection, is defined as a process of selecting a small subset of genes that contains the most discriminant information with well-defined evaluation metrics [6]. In addition to reducing the dimensionality of original gene space, effective gene selection methods bring us significant enhancements of quality measures for defining gene sets that validate the drug targets in biological and medical research. These enhancements include better generalization capacity of the constructed classifier, reducing the classifier training time, and improving the interpretability of obtained biomarkers [15].

According to whether using a classifier to evaluate the quality of a candidate feature in the feature selection process, existing feature selection methods can be broadly divided into four categories: (1) filter methods, (2) wrapper methods, (3) embedded methods, and (4) hybrid methods [16,17]. Filter methods are independent of a classification model and measure the quality of a feature, or a subset, using only the intrinsic nature of training samples. Filter methods are flexible in combination with various classifiers and have lower computational complexity. They also have better generalization ability [16]. Furthermore, commonly used metrics in filter methods mainly include distance, consistency, dependency, and information theory-based metrics [18]. Distance-based methods define separability as the metric and try to find those features that can best discriminate the target class. One such method is the ReliefF algorithm [19]. Consistency-based methods use the inconsistency rate as the criterion and seek to select a subset of features with better consistency, such as Focus and LVF algorithms [20]. Dependency-based methods evaluate the importance of candidate features with statistical theory, and there are a variety of methods available such as Pearson correlation coefficient, partial least squares, and Fisher score [21,22]. Information theory based feature selectors have efficiency and effectiveness because of their capacity in capturing higher order statistics of data and reflecting the non-linear relationships between variables [23]. Consequently, researchers have proposed and developed a number of feature selectors from the view of mutual information, including information gain, minimum redundancy maximum relevance (mRMR) [24], and fast correlation based filter (FCBF) [25]. In contrast, wrapper-based methods are specific to a given learning algorithm to extending non-filter features of selection to evaluate the quality of a selected candidate. These methods often use the classification error rate or classification accuracy as an evaluation criterion [26–28]. Due to the specific interaction between the obtained features and the learning algorithm, wrapper methods tend to obtain better classification results but at the cost of high time complexity [27]. Embedded methods are essentially a special case of wrapper methods and more tightly coupled with a specified learning algorithm. Feature subsets are generated during training the classifier, which makes embedded methods usually more tractable and time efficient than wrapper methods. Decision tree and Lasso algorithms are two typical embedded cases [29,30]. Besides, a hybrid scheme has been proposed to take advantage of both filter and wrapper methods, and researchers have proposed to combine filter and wrapper methods [31,32]. Essentially, a filter is initially used to remove a large number of noisy and redundant features from the original feature space, and then a wrapper method is used to find a discriminant feature subset from the reduced subset [33].

According to the final output style, we can group existing feature selection methods into feature ranking and feature subset selection categories. Feature ranking methods return a ranked list of the original features in descending order according to the predictive power of each feature [34]. We are required to specify the number of how many features are to be selected after ranking. Alternatively, we can determine the optimal size of a feature subset with the help of a learning algorithm. Feature ranking methods include single feature ranking and multiple feature ranking methods. The former evaluates the quality of each candidate feature individually, and does not consider the redundancy and interaction between features [19]. These feature ranking methods often fail to obtain a feature subset of high quality. Multiple feature ranking methods take the relationship between the candidate feature and previously selected features into account in the process of feature selection [25]. Ranking methods belonging to this category have a sequential forward or backward selection scheme to rank original features [6]. Unlike feature ranking methods, feature subset selection methods explicitly or implicitly consider the relevance and redundancy between features, and finally return a feature subset without involving a further step to determine the optimal size [25].

Currently, there are a wealth of feature selection methods available [35–38], but most of them seek to find a common subset of genes for subtypes within a defined pathology, and fail to reflect the unique characteristics of each subtype based on molecular differences. In fact, a unique subset of genes is likely to exist within each tumor subtype. Identifying these molecularly based tumor subtypes will increase the clinical efficacy of treatments with such predictive biomarkers [2,39]. Obtaining molecular subtype dependent biomarkers helps design a personalized treatment plan. These plans have been shown to often reduce the toxicity and side effects in treatment, concurrent with significant slowing of tumor progression. These biomarkers also accelerate structural and cell-based refinement in drug development research on these molecular subtypes, reducing the time and cost of bring drugs to clinic.

There are studies from related fields that propose to select a possible different feature subset for each class. For example, de Lannoy et al. propose a method to perform class-specific feature selection in multiclass support vector machines and experimentally validate its performance [40]. Zhou and Wang use class separability measure to select different feature subsets for different classes and compare their method with class independent feature selection method by applying the method on several biomedical data with support vector machine [41]. A major limitation of these methods is that they are related to the use of a particular classifier, which limits its applicability. To alleviate this problem, Pineda-Bautista et al. propose a class specific feature method that can be used with any classifiers and they use classifier ensemble to classify an unseen sample. Their experimental results on low dimensional datasets show the effectiveness of the proposed method [42]. However, classifying new test samples under an ensemble framework without utilizing the confidence of each sub-classifier may makes poor decisions when we face the problem of voting conflict. Besides, the aim of these studies is to return multiple feature subsets for feature analysis and classification model construction, and few studies, to the best of our knowledge, explore the fusion of multiple class-specific feature subsets and further evaluate the effectiveness of these combined features in classification. Furthermore, they conduct experiments on low dimensional datasets without considering a more difficult case that is characterized by high dimensionality and small sample sizes. Accordingly, in this study, we propose to select gene profiles that are associated tumor subtypes, enabling us to define unique genes for a tumor subtype as well as common genes for all tumor subtypes. We will enhance the performance in classifying different tumor subtypes and further reduce the chance of partially overfitting in future algorithms. The main contributions of this study are as follows:

Please cite this article as: A. Wang et al., Subtype dependent biomarker identification and tumor classification from gene expression profiles, Knowledge-Based Systems (2018), https://doi.org/10.1016/j.knosys.2018.01.025