Journal of Biomedical Informatics 46 (2013) 1044–1059

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A simulation to analyze feature selection methods utilizing gene ontology for gene expression classification

Christopher E. Gillies [a], Mohammad-Reza Siadat [a,*], Nilesh V. Patel [a], George D. Wilson [b]

[a] Dept. of Computer Science and Engineering, Oakland University, 2200 N Squirrel Rd, Rochester, MI 48309, United States
[b] Beaumont Health System, 3601 W. Thirteen Mile Rd, Royal Oak, MI 48073, United States

ARTICLE INFO

Article history:
Received 26 September 2012
Accepted 21 July 2013
Available online 25 July 2013

Keywords:
Data mining
Gene expression
Cancer classification
Gene ontology
Semantic similarity
Feature evaluation and selection

ABSTRACT

Gene expression profile classification is a pivotal research domain assisting in the transformation from traditional to personalized medicine. A major challenge associated with gene expression data classification is the small number of samples relative to the large number of genes. To address this problem, researchers have devised various feature selection algorithms to reduce the number of genes. Recent studies have been experimenting with the use of semantic similarity between genes in Gene Ontology (GO) as a method to improve feature selection. While there are few studies that discuss *how* to use GO for feature selection, there is no simulation study that addresses *when* to use GO-based feature selection. To investigate this, we developed a novel simulation, which generates binary class datasets, where the differentially expressed genes between two classes have some underlying relationship in GO. This allows us to investigate the effects of various factors such as the relative connectedness of the underlying genes in GO, the mean magnitude of separation between differentially expressed genes denoted by δ, and the number of training samples. Our simulation results suggest that the connectedness in GO of the differentially expressed genes for a biological condition is the primary factor for determining the efficacy of GO-based feature selection. In particular, as the connectedness of differentially expressed genes increases, the classification accuracy improvement increases. To quantify this notion of connectedness, we defined a measure called Biological Condition Annotation Level $BCAL(G)$, where G is a graph of differentially expressed genes. Our main conclusions with respect to GO-based feature selection are the following: (1) it increases classification accuracy when $BCAL(G) \geqslant 0.696$; (2) it decreases classification accuracy when $BCAL(G) \leqslant 0.389$; (3) it provides marginal accuracy improvement when $0.389 < BCAL(G) < 0.696$ and $\delta < 1$; (4) as the number of genes in a biological condition increases beyond 50 and $\delta \geqslant 0.7$, the improvement from GO-based feature selection decreases; and (5) we recommend not using GO-based feature selection when a biological condition has less than ten genes. Our results are derived from datasets preprocessed using RMA (Robust Multi-array Average), cases where δ is between 0.3 and 2.5, and training sample sizes between 20 and 200, therefore our conclusions are limited to these specifications. Overall, this simulation is innovative and addresses the question of *when* SoFoCles-style feature selection should be used for classification instead of statistical-based ranking measures.

1. Introduction

A major transformation is occurring within the health-care community. Instead of applying general treatments to diverse patient populations, therapies are being customized to patient subpopulations based on their gene expression profiles. This new form of medical practice is known as personalized medicine. Gene expression profile classification is subfield of bioinformatics that is aiding in the transformation to personalized medicine. Gene expression profiling is an important tool for personalized medicine because it allows biomedical researchers to discover biomarkers. There are two types of biomarkers, prognostic biomarkers and predictive biomarkers. Prognostic biomarkers allow clinicians to discern which patients to treat, while predictive biomarkers elucidate a treatment's effectiveness for a patient [1]. For gene expression studies, a biomarker is represented by the expression of a gene or set of genes under a certain physiological condition. To give the reader some insight regarding biomarkers, suppose some treatment t_{reat} has been shown to be effective for patients having some biomarker b in physiological situation s_{it}. Now suppose we have a patient p with physiological situation s_{it}, we want to say: if patient p with physiological situation s_{it} has biomarker b then apply treatment t_{reat} to patient p. In this example, b is a

* Corresponding author.
E-mail addresses: cegillie@oakland.edu (C.E. Gillies), siadat@oakland.edu (M.-R. Siadat), npatel@oakland.edu (N.V. Patel), george.wilson@beaumont.edu (G.D. Wilson).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.07.008

predictive biomarker. There are many applications of gene expression profile classification including: tumor class discrimination, prediction of clinical outcome based on treatments and detection of previously unknown sub-patterns [2]. Gene expression profiles have been traditionally collected using DNA microarrays, however, recent studies are also using RNA-seq [3]. In its simplest form, the gene expression classification problem compares two classes: (1) a control class and (2) an experimental class. Dubitzky et al. describe nine steps involved with microarray data analysis: (1) identify scientific aims; (2) design experiment; (3) design/make or acquire microarray; (4) hybridize and scan microarray; (5) analyze resulting image; (6) derive data matrix; (7) pre-process data matrix; (8) analyze and model; and (9) interpret and validate results [4].

In this paper, we focus on steps seven and eight of microarray data analysis. Step seven has a few subtasks such as missing value computation, normalization, transformation and feature selection. Within step seven, we are most interested in feature selection. The ultimate goal of feature selection, also known as gene selection, is to reduce the dimensionality of the problem and identify potential biomarkers. Feature selection is supremely important because a gene expression profile has thousands of values associated with it, and fitting a classifier with exceptionally high dimension leads to the curse of dimensionality. Adding to the problem is the fact that there are usually only tens or hundreds of gene expression profiles to use as training examples. With regards to step eight, many different classifiers such as linear discriminant analysis (LDA) [5], diagonal linear discriminant analysis (DLDA) [5], weighted voting (WV) [6], k-nearest neighbor (KNN) [7] and, support vector machines (SVM) [8] have been applied on gene expression profiles [9].

To frame the problem in a more mathematical context, let's define the training set to be $T \in \mathbb{R}^{m \times n}$ where m is the number of genes, n is the number of biospecimens analyzed and $\mathbb{R}^{m \times n}$ refers to a m by n dimensional real number space. A column of the set T represents the gene expression profile of a biospecimen; a row represents the expression levels for a single gene across all biospecimens. The primary objective of feature selection is to find $T' \in \mathbb{R}^{d \times n}$ where $T' \subset T$ and $d < m$ such that training a classifier C on T' yields higher generalized classification accuracy than training on T.

Given the importance of discovering biomarkers, one should not be surprised to find a vast amount of literature devising feature selection algorithms. There are three common approaches to feature selection: filter, wrapper, and embedded techniques [10]. To understand how feature selection algorithms are classified using this scheme, it is useful to envision these processes as they occur along a time-line with respect to training a classifier. Specifically, filtering occurs before classifier training, embedded selection occurs during classifier training and wrappers are applied after classifier training. Filtering techniques typically rank genes by some statistical metric and then remove all genes that fall below a user-defined threshold. Wrapper methods attempt to find an optimal subset of genes that achieve high accuracy. These methods are called wrappers because they encapsulate a classifier and call the classifier as a subroutine. Table 1 lists some feature selection techniques.

Table 1
Examples of feature selection techniques.

Technique	Type	Publication
Signal-to-noise ratio	Filter	[6]
t-Statistics	Filter	[11]
ANOVA	Filter	[12]
Wilcoxon rank-sum	Filter	[13]
BLOCK.FS	Wrapper	[14]
Multiple SVM-RFE	Wrapper	[15]
Integer-coded genetic algorithm	Wrapper	[16]
Genetic programming	Embedded	[17]
Multiple-filter–multiple-wrapper	Combination	[9]

The previously mentioned techniques discover important genes by comparing statistical properties of a dataset, and do not include domain specific biological knowledge into the selection. Recently, some researchers have been investigating whether or not prior knowledge could improve feature selection for gene expression data classification. The rationale for these investigations is based on the following inductive argument: (1) gene expression data has small sample sizes, so the identification of important genes is difficult; (2) there are large biomedical knowledge-bases such as Gene Ontology (GO) [18] and Gene Ontology Annotation (GOA) [19] that describe gene relationships; (3) there appears to be some correlation between gene expression data and semantic similarity between terms in GO [20,21]; (4) there has been success by incorporating prior knowledge in other pattern recognition tasks; therefore, it seems possible that incorporating prior knowledge into feature selection techniques will improve biomarker identification and classification accuracy for gene expression data.

Further support for feature selection techniques that incorporate prior knowledge can be found in the success of enrichment analysis tools. The purpose of enrichment analysis tools is to assist with the interpretation of a list of relevant genes from data generated using high-throughput technologies like microarrays. Huang et al. mention that these tools are built on the following assumption: if a biological process is not functioning properly, then genes involved in this biological process will have a higher likelihood to be relevant [22]. The goal of these enrichment analysis tools is to find biological processes that best describe a user-specified list of relevant genes. Some of these enrichment analysis tools discover relevant biological terms by comparing a GO term's coverage among the list of relevant genes to its coverage among all genes. The difference between feature selection methods and enrichment analysis tools is that feature selection methods build a list of relevant genes, where as enrichment analysis tools assist with the interpretation of a list of relevant genes.

Some examples of enrichment analysis tools are Onto-Express [23], MaPPFinder [24], GOMiner [25], DAVID [26], EASE [27], GeneMerge [28], and FuncAssociate [29]. Refinements to enrichment analysis tools using information theory can be found in [30]. Other enrichment analysis tools, do not require a list of relevant genes, instead they work on all the genes. An example is Gene Set Enrichment Analysis (GSEA) discussed in [31,32]. GSEA works on a ranked list of genes that are correlated with a phenotype. GSEA tries to discover functional annotations such as GO terms that are either up-regulated or down-regulated relative to a control group. This allows for functional annotation-level analysis.

Extensions of functional annotation-level analysis methods are found in signatures of pathway deregulation in tumors [33], Condition-Responsive Genes (CORGs) [34] and the Functional Analysis of Individual Microarray Expression (FAIME) profiles [35]. Two other functional-level analysis methods aimed at the interpretation of high-throughput biological results are [36,37]. Functional-level analysis, similar to other enrichment analysis tools, are also used to interpret high-throughput results, however, methods like FAIME map gene expression values onto functional-level annotations such as GO terms. This mapping procedure allows pattern recognition tasks to be performed directly at the functional-level instead of at the gene-level. Feature selection methods, as discussed in this paper, select important genes at the gene-level.

SoFoCles [38] is a feature selection technique, which is based, in part, on Qi and Tang's method [39,40]. SoFoCles uses information from GO to improve statistical feature selection. In our paper, we refer to the enrichment of feature selection using GO as GO-based feature selection. The authors of SoFoCles show that GO-based