Journal of Biomedical Informatics 64 (2016) 255–264

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Special Communication

Sensitivity analysis of gene ranking methods in phenotype prediction

Enrique J. deAndrés-Galiana [a,b], Juan L. Fernández-Martínez [a,*], Stephen T. Sonis [c]

[a] Department of Mathematics, University of Oviedo, Spain
[b] Artificial Intelligence Center, University of Oviedo, Spain
[c] Biomodels, LLC, Watertown, MA, USA

ARTICLE INFO

Article history:
Received 12 February 2016
Revised 20 September 2016
Accepted 24 October 2016
Available online 26 October 2016

Keywords:
Noise analysis
Machine learning
Gene expression
Cancer genomics

ABSTRACT

Introduction: It has become clear that noise generated during the assay and analytical processes has the ability to disrupt accurate interpretation of genomic studies. Not only does such noise impact the scientific validity and costs of studies, but when assessed in the context of clinically translatable indications such as phenotype prediction, it can lead to inaccurate conclusions that could ultimately impact patients. We applied a sequence of ranking methods to damp noise associated with microarray outputs, and then tested the utility of the approach in three disease indications using publically available datasets.

Materials and methods: This study was performed in three phases. We first theoretically analyzed the effect of noise in phenotype prediction problems showing that it can be expressed as a modeling error that partially falsifies the pathways. Secondly, via synthetic modeling, we performed the sensitivity analysis for the main gene ranking methods to different types of noise. Finally, we studied the predictive accuracy of the gene lists provided by these ranking methods in synthetic data and in three different datasets related to cancer, rare and neurodegenerative diseases to better understand the translational aspects of our findings.

Results and discussion: In the case of synthetic modeling, we showed that Fisher's Ratio (FR) was the most robust gene ranking method in terms of precision for all the types of noise at different levels. Significance Analysis of Microarrays (SAM) provided slightly lower performance and the rest of the methods (fold change, entropy and maximum percentile distance) were much less precise and accurate. The predictive accuracy of the smallest set of high discriminatory probes was similar for all the methods in the case of Gaussian and Log-Gaussian noise. In the case of class assignment noise, the predictive accuracy of SAM and FR is higher. Finally, for real datasets (Chronic Lymphocytic Leukemia, Inclusion Body Myositis and Amyotrophic Lateral Sclerosis) we found that FR and SAM provided the highest predictive accuracies with the smallest number of genes. Biological pathways were found with an expanded list of genes whose discriminatory power has been established via FR.

Conclusions: We have shown that noise in expression data and class assignment partially falsifies the sets of discriminatory probes in phenotype prediction problems. FR and SAM better exploit the principle of parsimony and are able to find subsets with less number of high discriminatory genes. The predictive accuracy and the precision are two different metrics to select the important genes, since in the presence of noise the most predictive genes do not completely coincide with those that are related to the phenotype. Based on the synthetic results, FR and SAM are recommended to unravel the biological pathways that are involved in the disease development.

1. Introduction

The revolution in molecular biology and the development of high-throughput technologies for sequencing in genetic and genomic analyses has generated an explosion in the amount of genetic data. These technologies, which have been firstly applied in research, are now increasingly applied in translational medicine. Particularly, gene expression analysis through hybridization microarrays or RNA sequencing is now a conventional component in many areas of biomedical research. This kind of experiments has a very high under-determined character since the number of samples (patients) is much lower than the number of monitored probes (genes). Therefore, gene-ranking methods are needed to establish the discriminatory power of the genes in the phenotype prediction.

* Corresponding author.
E-mail address: jlfm@uniovi.es (J.L. Fernández-Martínez).

http://dx.doi.org/10.1016/j.jbi.2016.10.012
1532-0464/

In this paper we first theoretically analyzed the effect of noise in phenotype prediction problems by casting them into abstract optimization problems. To accomplish this, we first show that noise in data can be expressed as a modeling error that partially falsifies the set of discriminatory probes that are phenotype-related, and therefore the biological pathways that are involved. Secondly, the sensitivity to different kind of noise (in expression and class assignment) for the main gene ranking methods (Fold Change, Fisher's Ratio, Percentile Distance and Entropy) compared to well-established Significance Analysis of Microarrays (SAM) [1] is performed via synthetic microarray modeling. This analysis has shown that in general terms Fisher's ratio is the most robust method in terms of precision closely followed by SAM. Besides, both methods provided the smallest sets with the highest discriminatory power. The effect of noise increases the number of genetic probes that are needed to slightly improve the predictive accuracy. Therefore, an optimum method to find the biological pathways in translational problems will consist of ranking the differential expressed genes decreasingly by their corresponding Fisher's ratio. The results of these analyses are confirmed using three different datasets concerning the study of cancer (Chronic Lymphocytic Leukemia), rare diseases (Inclusion Body Myositis) and neurodegenerative diseases (Amyotrophic Lateral Sclerosis). We found that FR and SAM provide the highest predictive accuracies with the smallest number of genes, exploiting the principle of parsimony. Besides, we show their corresponding biological found with an expanded list of genes whose discriminatory power has been established via FR. In these three cases, the effect of viral infections in the corresponding pathways is clear. We demonstrated that applying a proper ranking method the influence of noise in microarray expression dataset and the corresponding error in the classification induced by the different sources of noise can be reduced. Similarly to the analysis shown in this paper, Lorena et al. [2] have studied the particular characteristics of cancer gene expression data mostly impact the prediction ability of support vector machine classifiers. We expect that the results of this analysis will help optimize the use of these methods in translational medicine, particularly in the biological understanding of different diseases and in drug optimization problems.

2. Material and methods

2.1. The effect of noise in phenotype prediction

One of the main obstacles in the analysis of genomic data is the absence of a conceptual model that relates the different genes/probes to the class prediction (phenotype). Machine-learning algorithms are therefore needed to model these complex relationships. For this reason, a classifier $L^*(\mathbf{g})$ has to be constructed and it is defined as an application between the set of genetic signatures $\mathbf{g}$ and the set of classes $C = \{c_1, c_2, \ldots, c_n\}$ in which the phenotype is divided:

$$L^*(\mathbf{g}) : \mathbf{g} \in \mathbf{R}^s \rightarrow C = \{c_1, c_2, \ldots, c_n\} \tag{1}$$

where s is the number of genetic probes that have been monitored.

In most cases, the classification problem involved is binary.

The machine learning procedure is composed of two stages:

1. The learning process, that consists in giving a subset of samples $\mathbf{T}$ (training data set) whose class vector is known, $\mathbf{c}^{obs}$, finding the subset of genetic signatures $\mathbf{g}$ that maximizes the learning accuracy, that is, the number of samples whose class is correctly predicted. This can be written as the result of the following optimization problem:

$$\begin{aligned} O(\widetilde{\mathbf{g}}) &= \min_{\mathbf{g}\in\mathbf{R}^s} O(\mathbf{g}), \\ O(\mathbf{g}) &= \|\mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{obs}\|_p, \\ \mathbf{L}^*(\mathbf{g}) &= (L^*(\mathbf{g}_1), \ldots, L^*(\mathbf{g}_m)), \end{aligned} \tag{2}$$

where $\mathbf{L}^*(\mathbf{g})$ is the set of predicted classes, $\mathbf{g}_i$ is the genetic signature corresponding to the sample i in the training dataset $\mathbf{T}$, and $\|\mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{obs}\|_p$ stands for the distance between the predicted ($\mathbf{L}^*(\mathbf{g})$) and observed classes $\mathbf{c}^{obs}$ in $\mathbf{T}$. For instance if the vector of classes $\mathbf{c}^{obs}$ is composed of two consecutive class indexes $\{1,2\}$, and $p = 1$, then $\|\mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{obs}\|_p$ provides the number of samples that have been misclassified. Nevertheless, in this paper the theoretical analysis is performed for any arbitrary norm, in order to understand the impact of noise in phenotype prediction.

2. The generalization, that consists in predicting the class of a new sample ($\mathbf{g}_{new}$) whose class is unknown using the genetic signatures that have been found during the learning process.

One of the main numerical difficulties in learning is the high dimensionality of the genomic data since the number of monitored probes (or genes) is much greater than the number of samples (or patients). This fact provokes that the phenotype prediction in the learning stage will have a very high underdetermined character. Therefore, several gene lists with similar predictive accuracy might exist. This fact can be easily understood considering the classification as a parameter identification or inverse problem [3]: the topography of the cost function $O(\mathbf{g})$ in the region of lower misfits (or higher predictive accuracies) corresponds to flat elongated valleys with null gradients where the high predictive genetic signatures are located. Obviously, the topography changes if the space where the optimization is performed ($\mathbf{R}^s$) changes. All these high predictive lists are expected to be involved in the genetic pathways that explain the phenotype. The smallest-scale signature is the one that has the least number of discriminatory genes. In practice, the predictive accuracy of a genetic signature, $O(\mathbf{g})$, is performed via cross-validation. This knowledge could be very important for early diagnosis and treatment optimization. Also, the sets of high predictive signatures in any phenotype prediction problem can be used to construct biomedical robots that exploit the uncertainty space of the phenotype prediction, to improve the predictive accuracy of the classifier with its corresponding risk assessment, helping to get a better understanding the biological pathways [4].

The presence of noise in the genomic data will impact the classification and obviously the pathway analysis resulting from this procedure. There are at least two main sources of noise in phenotype prediction problems:

- **Noise in the gene expression** induced by the process of measurement. In this case, the observed genetic expression of a sample j, $\mathbf{g}_j^{obs}$, can be expressed as the sum of the true genetic expression array, $\mathbf{g}_j^{true}$, and the measurement noise, $\delta\mathbf{g}_j : \mathbf{g}_j^{obs} = \mathbf{g}_j^{true} + \delta\mathbf{g}_j$. Therefore, using a simple Taylor expansion we get:

$$\begin{aligned} L^*(\mathbf{g}_j^{obs}) &= L^*(\mathbf{g}_j^{true}) + \delta L^*(\delta\mathbf{g}_j) \\ &= L^*(\mathbf{g}_j^{true}) + \sum_{k=1}^{s} \frac{\partial L^*}{\partial g_k}(\mathbf{g}_j^{true})\delta g_{jk} + o(\delta\mathbf{g}_j), \end{aligned} \tag{3}$$

where $o(\delta\mathbf{g}_j)$ vanishes when the noise term $\delta\mathbf{g}_j \rightarrow \mathbf{0}$. Obviously, this analysis is theoretical because $\mathbf{g}_j^{true}$ and $\delta\mathbf{g}_j$ are unknown.

- **Noise in the class assignment** since some samples could be wrongly annotated or might belong to a different class, not yet discovered. Naming $\mathbf{c}^{true}$ the true class assignment array and $\delta\mathbf{c}$ the noise in the class assignment, then the observed class array will be $\mathbf{c}^{obs} = \mathbf{c}^{true} + \delta\mathbf{c}$.