COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

Volume No: 6, Issue: 7, March 2013, e201303004, http://dx.doi.org/10.5936/csbj.201303004

Biomarker signature identification in "omics" data with multi-class outcome

Vincenzo Lagani [a,*], George Kortas [b], Ioannis Tsamardinos [a,b]

Abstract: Biomarker signature identification in "omics" data is a complex challenge that requires specialized feature selection algorithms. The objective of these algorithms is to select the smallest set(s) of molecular quantities that are able to predict a given outcome (target) with maximal predictive performance. This task is even more challenging when the outcome comprises of multiple classes; for example, one may be interested in identifying the genes whose expressions allow discrimination among different types of cancer (nominal outcome) or among different stages of the same cancer, e.g. Stage 1, 2, 3 and 4 of Lung Adenocarcinoma (ordinal outcome). In this work, we consider a particular type of successful feature selection methods, named constraint-based, local causal discovery algorithms. These algorithms depend on performing a series of conditional independence tests. We extend these algorithms for the analysis of problems with continuous predictors and multi-class outcomes, by developing and equipping them with an appropriate conditional independence test procedure for both nominal and ordinal multi-class targets. The test is based on *multinomial logistic* regression and employs the log-likelihood ratio test for model selection. We present a comparative, experimental evaluation on seven real-world, high-dimensional, gene-expression datasets. Within the scope of our analysis the results indicate that the new conditional independence test allows the identification of smaller and better performing signatures for multi-class outcome datasets, with respect to the current alternatives for performing the independence tests.

7TH CONFERENCE OF THE HELLENIC SOCIETY FOR COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

Introduction

Microarray technologies and Next Generation Sequencing (NGS) techniques nowadays allow precise measurements of several types of within-cell molecular quantities. Gene expression data, methylation level measurements, proteomics and metabolomics information are just a few example of the "omics" data that such technologies are able to provide.

Even though "omics" technologies simultaneously measure tens of thousands or more molecular quantities, researchers are often interested in identifying a relatively small subset of such measurements, known as *biomarkers,* which are relevant for the problem under study. A typical example is the identification of genes whose expressions statistically significantly differ among two or more conditions (i.e., differentially expressed genes).

Identifying biomarkers that are relevant (informative) when considered *in isolation* is often useful for investigating biological systems' underlying mechanisms. However, in some cases the identification of *biomarker signatures* is necessary instead. We define a biomarker signature as a minimal subset of molecular quantities that are maximally informative for a given task when considered jointly.

For example, a researcher may be interested in finding the smallest subset of genetic variants (e.g., Single Nucleotide Polymorphisms, SNPs) that considered together are maximally predictive with respect to the development of osteoporosis in elderly women. As a further example, one may focus on finding the smallest number of CpG sites whose methylation levels discriminate with the maximal possible accuracy among different types of lung cancers. In both examples it is crucial to identify the set of biomarkers providing the highest possible accuracy; at the same time, it is necessary to take into account that the cost of devising, realizing and routinely performing a clinical test based on the signature is probably directly related to the number of involved biomarkers.

A widely adopted approach for identifying new biomarker signatures consists in measuring a large set of molecular quantities from a sufficiently large sample of biological specimens, and then employing data-analysis approaches in order to select the most informative set of features.

In the fields of statistics and machine learning the task of identifying the most relevant variables for the problem at hand is known as *feature* or *variable selection* [1]. Numerous methods have been developed for addressing the problem. A recent successful approach based on local causal discovery, namely the Max-Min Parent Children (MMPC) algorithm, has been described in [2]. Unlike some other methods, this approach is principled in the sense that it provides theoretical guarantees under which the methods soundly solve the feature selection problem. In particular, MMPC attempts to retrieve (a subset of) the *Markov-Blanket* (MB) of the considered outcome. The MB of a target variable is the set of variables conditioned upon which any other set of variables becomes independent by the target. It has been theoretically demonstrated that, under broad assumptions, the MB of a variable coincides with its minimal-size, most-informative signature [3]. This theoretical result was recently

[a]Institute of Computer Science, Foundation for Research and Technology – Hellas (FORTH), N. Plastira 100, Vassilika Vouton, GR-700 13 Heraklion, Crete, Greece

[b]Department of Computer Science, University of Crete, P.O.Box 2208, GR-710 03 Heraklion, Crete, Greece

* Corresponding author. Tel.: +30 2810391070; Fax: +30 2810391428
E-mail address: vlagani@ics.forth.gr (Vincenzo Lagani)

supported by large scale evaluations [4,5] that have experimentally demonstrated local causal discovery algorithms' efficacy in finding highly predictive signatures.

MMPC operates as a Constraint-Based (CB) variable selection algorithm. The operation of all such CB methods iteratively applies, based on a search strategy, *Conditional Independence Tests* (CITs) for characterizing the data distribution and identifying the variables (not) belonging to the MB. CITs, hereafter represented as Test(X, Y |**Z**), are statistical procedures that assess the null hypothesis "X and Y are independent given **Z**", where X and Y are two random variables, the conditioning set **Z** is a (possibly empty) set of random variables, and X,Y $\notin$ **Z**. Intuitively, a CIT assesses whether X gives any additional information for Y (and vice-versa) once **Z** is known.

CITs role within CB algorithms is pivotal; employing an inappropriate CIT would lead to a poorly approximated MB and consequently to a suboptimal signature. The Fisher Z test [6] is currently the most widely employed conditional independence test for cases when all variables, including the target, are continuous. The Fisher test assumes linear relations among variables as well as normally distributed error terms: assumptions that are quite unlikely to hold in omics data. For discrete data testing is typically implemented with asymptotic G^2 and χ^2 tests [7] or exact permutation-based versions of these tests [8]. Attempts have been performed in order to develop sample-efficient CIT not relying on parametric assumptions [9], but further research in this field is needed: in particular, large scale evaluations for comparing different CITs' respective performances are largely missing. To the best of our knowledge, up to date no CIT for continuous predictors / multi-class target has been applied and evaluated for CB algorithms.

In this paper, we devise a CIT specifically for cases where all predictors are numerical (continuous) and the target outcome represents multiple classes (categories). This is a common scenario in studies dealing with "omics" data that look for molecular signatures able to discriminate among different conditions (e.g., different cancer stages or types). The Fisher Z test is usually employed in these settings, after encoding the outcome as a discrete, integer variable; this workaround introduces a possibly unnatural order among outcome categories and assumes linear relationships among regressors and outcome. The CIT we develop, named Multi-Class Conditional Independence Test (MC-CIT) is based on the multinomial logistic regression and is turned into a test by employing the log-likelihood ratio test for model selection [10]. The multinomial Logistic models are Generalized Linear Models (GLM [11]) specifically devised for modeling multi-class outcomes; we thus expect MC-CIT to outperform the Fisher Z test in such settings.

In order to support our claim, we contrasted the newly proposed test against both the prototypical Fisher Z and G^2 tests, in an extensive evaluation involving seven high-dimensional, multi-class gene expression studies. The following sections describe MC-CIT theoretical basis and implementation details, along with the experimentation protocol employed for assessing its performances.

Notably, the results of the experimentation underlined the superior performances of MC-CIT against the Fisher Z test, in terms of both predictive capability and parsimoniousness of the selected biomarker signatures.

Experimental procedures

MC-CIT: Conditional Independence Test based on Multinomial Logistic Regression

Let's assume that Y is a categorical random variable representing a multi-class outcome, X a continuous random variable and **Z** a set of continuous random variables; let's indicate with *Ind*(X,Y|**Z**) the MC-CIT null hypothesis of independence "Y is independent by X given **Z**". Assuming that *Ind*(X,Y|**Z**) holds implies that X is not necessary for predicting Y once **Z** is known; under this respect the MC-CIT can be devised as a nested-model selection procedure, where the "full" model employs {X, **Z**} as regressors for Y, while the alternative model employs only {**Z**} [12]. When the full model shows a statistically significantly better fit than the alternative model, then the null hypothesis *Ind*(X,Y|**Z**) can be rejected, i.e., X and Y are *associated* given **Z**. When the two models are statistically indistinguishable then the null hypothesis cannot be rejected and the algorithm accepts that the conditional independence holds.

Following these considerations, we implemented the MC-CIT as a log-likelihood ratio test for nested-model selection, where the full and alternative models are fitted with either a Multinomial Logistic (ML, for categorical outcomes) or with an Ordered Logit (OL [11], for ordered outcomes) regression approach. In both cases, let $LogL_{Full}$ and $LogL_{Altern}$ be the log-likelihood of the full and alternative model, respectively; then the quantity D

$$D = -2 \cdot (LogL_{Full} - LogL_{Altern})$$

follows a χ^2 distribution with one degree of freedom, under the assumption that *Ind*(X,Y|**Z**) holds. Given D and its theoretical distribution we calculate a p-value for the MC-CIT null hypothesis.

The key reason for preferring the ML and OL regressions over the simpler Fisher Z test linear approach is that these two regression procedures are specifically devised in order to model discrete outcomes over continuous regressors. We thus expect ML and OL to better model multi-outcome data and consequently to enhance the detection of true conditional dependencies. Specifically, given a set of N regressors **W** and a binary outcome T, the standard logistic regression model can be expressed as:

$$ln\left(\frac{Pr(T_i = 1)}{1 - Pr(T_i = 1)}\right) = \boldsymbol{\beta} \cdot \boldsymbol{W}_i$$

where $\boldsymbol{\beta}$ is the set of model's coefficients, $\mathbf{W}_i$ represents the values that the regressors assume for the i^{th} sample, and $Pr(T_i = 1)$ is the probability that $T_i = 1$.

The ML regression extends the standard logistic regression to categorical outcomes by introducing K-1 set of coefficient $\boldsymbol{\beta}_k$, where K is the number of different values that the outcome can assume:

$$ln\left(\frac{Pr(T_i = k)}{Pr(T_i = K)}\right) = \boldsymbol{\beta}_k \cdot \boldsymbol{W}_i, \qquad k = 1, \cdots, K-1$$

Few simple mathematical operations bring to the following formulation:

$$Pr(T_i = k) = Pr(T_i = K) \cdot e^{\boldsymbol{\beta}_k \cdot \boldsymbol{W}_i}, \qquad k = 1, \cdots, K-1$$

The probability $Pr(T_i = K)$ can be calculated by taking in consideration that the whole set of probabilities $Pr(T_i = k)$, k = 1, …, K must sum up to 1:

$$Pr(T_i = K) = \frac{1}{1 - \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \boldsymbol{W}_i}}$$