Available online at www.sciencedirect.com
ScienceDirect
Biomolecular Engineering 24 (2007) 273–282
Biomolecular Engineering
www.elsevier.com/locate/geneanabioeng

Improving cluster-based missing value estimation of DNA microarray data

Lígia P. Brás, José C. Menezes *

Centre for Chemical & Biological Engineering, Department of Chemical and Biological Engineering, IST, Technical University of Lisbon, Av. Rovisco Pais, P-1049-001 Lisbon, Portugal

Received 27 October 2006; received in revised form 21 February 2007; accepted 12 April 2007

Abstract

We present a modification of the weighted *K*-nearest neighbours imputation method (KNNimpute) for missing values (MVs) estimation in microarray data based on the reuse of estimated data. The method was called iterative KNN imputation (IKNNimpute) as the estimation is performed iteratively using the recently estimated values.

The estimation efficiency of IKNNimpute was assessed under different conditions (data type, fraction and structure of missing data) by the normalized root mean squared error (NRMSE) and the correlation coefficients between estimated and true values, and compared with that of other cluster-based estimation methods (KNNimpute and sequential KNN). We further investigated the influence of imputation on the detection of differentially expressed genes using SAM by examining the differentially expressed genes that are lost after MV estimation.

The performance measures give consistent results, indicating that the iterative procedure of IKNNimpute can enhance the prediction ability of cluster-based methods in the presence of high missing rates, in non-time series experiments and in data sets comprising both time series and non-time series data, because the information of the genes having MVs is used more efficiently and the iterative procedure allows refining the MV estimates. More importantly, IKNN has a smaller detrimental effect on the detection of differentially expressed genes.

Keywords: Missing value estimation; *K*-nearest neighbours; Gene expression data; DNA microarray data

1. Introduction

DNA microarrays are a high-throughput technology that allows for the simultaneously monitoring of the mRNA levels of thousands of genes in particular cells or tissues, giving a global view of gene expression (Lockhart and Winzeler, 2000; Schena et al., 1995; Schulze and Downward, 2001).

The data generated in a set of microarray experiments are usually gathered in a matrix with genes in rows and experimental conditions in columns. Frequently, these matrices contain missing values (MVs). This is due to the occurrence of imperfections during the microarray experiment (e.g. insufficient resolution, spotting problems, deposition of dust or scratches on the slide, hybridization failures) that create suspect values, which are usually thrown away and set as missing (Alizadeh et al., 2000). The in situ synthesized Affymetrix GeneChips and the spotted cDNA (or oligonucleotide) microarrays are the two most commonly used types of microarray technology. The redundancy in design used in a GeneChip (i.e. a gene is represented by a set of approximately 20 probe pairs) prevents the existence of MVs. This is not the case for spotted cDNA microarrays, where usually each spot is assigned to a unique gene, and the use of double to quadruple spots for a gene is currently an exception. So, the loss at a spot usually leads to the loss of information for a gene, and thus to a MV in the gene expression data matrix. Therefore, in this work we consider the estimation of MVs in gene expression data obtained from spotted cDNA microarrays.

In some microarray data sets, the proportion of MVs is significant. For example, some authors reported that the percentage of gene profiles with at least one MV can be higher than 85% (de Brevern et al., 2004). The presence of missing gene expression values constitutes a problem for downstream data analyses, since many of the methods employed (e.g. classification and model-based clustering techniques) require

* Corresponding author. Tel.: +351 218 417 347; fax: +351 218 419 197.
E-mail addresses: ligia.bras@ist.utl.pt (L.P. Brás), bsel@ist.utl.pt (J.C. Menezes).

1389-0344/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.bioeng.2007.04.003

complete matrices. Due to economic reasons or biological sample availability, repeating the microarray experiments in order to obtain a complete gene expression matrix is usually unfeasible, so other alternatives have to be considered. The simple approaches usually applied to handle missing gene expression entries include removing the genes with MVs before the analysis (case deletion), or replacing the MVs of a gene with the average of the observed values over that gene (mean substitution; Schafer and Graham, 2002). Another common approach is to replace missing $\log_2$ transformed gene expression ratios by zeros (Alizadeh et al., 2000). These approaches have disadvantages: case deletion procedures may bias the results if the remaining cases are unrepresentative of the entire sample (Little and Rubin, 1987), while both mean and zero substitutions distort relationships among variables and artificially reduce the variance of the variable in question (Little and Rubin, 1987; Schafer and Graham, 2002), since the same value is used to replace missing entries in a given gene.

To overcome these drawbacks, Troyanskaya et al. (2001) proposed a method called weighted *K*-nearest neighbour imputation (KNNimpute) that reconstructs the MVs using a weighted average of *K* most similar genes. Overall, this estimation method is more robust than others, such as replacement by zero, row average or singular value decomposition, to the fraction of missing elements and to the type of data for which estimation is executed, performing better in non-time series data or noisy data (Troyanskaya et al., 2001). As an improvement of KNN imputation, Kim et al. (2004) proposed a sequential KNN imputation method (SKNNimpute) that uses the estimated values sequentially for the later nearest neighbour calculation and estimation.

In a recent work, de Brevern et al. (2004) studied the stability of gene clusters of microarray data including MVs or not, specified by diverse hierarchical clustering algorithms, showing that the MVs (even at a low rate) have important effects on the gene clusters' stability. Thus, the presence of MVs in the data matrix should not be neglected, and MV estimation should be regarded as a pre-processing step essential to obtain proper results from microarray data analyses.

Although other methods have been proposed for estimating gene expression missing data, such as regression-based methods (Bø et al., 2004; Kim et al., 2005; Nguyen et al., 2004; Brás and Menezes, 2006) and Bayesian approaches (Oba et al., 2003), in this work we focus on the cluster-based methods, since these are widely used for the replacement of MVs in microarray data. For example, KNNimpute is the only imputation method available in significance analysis of microarrays (SAM; Tusher et al., 2001), prediction analysis for microarrays (PAM; Tibshirani et al., 2002) and microarray analysis of variance (MAANOVA; Kerr et al., 2000).

We propose an iterative procedure for the prediction of gene expression MVs called iterative KNN imputation (IKNNimpute), and compare its performance with that of other clustering-based imputation methods (KNNimpute and SKNNimpute) for various rates of MVs and type of missing structure using publicly available microarray data sets.

The methods are evaluated by comparing their estimates for the artificial missing entries with the true values, using measures such as normalized root mean squared errors, correlation coefficients and bias. Though such approach gives important measures of performance, a more fundamental and functional question that should further be addressed is the effect of the methods' estimates on the final output of different analysis methods, such as clustering algorithms or statistical algorithms for the differential analysis of gene expression. In the literature, such evaluations are lacking, and only a few cases can be found (for example, see de Brevern et al., 2004; Ouyang et al., 2004; Scheel et al., 2005; Jörnsten et al., 2005). In our study, the impact of the imputation methods' estimates on significance analysis for differential expression is also performed by comparing the lists of differentially expressed genes obtained using the statistical method known as SAM (Tusher et al., 2001). We opted to focus on the effects of imputation on differential expression, since, although cluster analysis of microarray data is capable of discovering coherent patterns of gene expression, it gives little information about statistical significance, i.e., about whether changes in gene expression are experimentally significant.

2. Materials and methods

2.1. Notation

Throughout this paper, microarray data are represented by matrices with rows corresponding to genes and columns to experimental conditions. In particular, $\boldsymbol{G}$ represents the original data matrix (with real MVs), while $\boldsymbol{X}$ is a gene expression matrix with p genes and n experiments (with $p \gg n$) that may contain missing entries. The ith row of $\boldsymbol{X}$ represents the expression profile of the ith gene in the n experiments, whereas x_{ij} denotes the expression level of gene i in sample j.

Using the notation of Nguyen et al. (2004), a gene with MVs is called *target* gene, and the genes with available information for estimating its missing entries constitute the set of *candidate* genes.

We also make use of the missing indicator matrix, $\boldsymbol{R}$, defined by Rubin (1976) to track the missing and non-missing entries of $\boldsymbol{X}$. If the expression value x_{ij} is available, the ijth element of $\boldsymbol{R}$, r_{ij}, is equal to 1, otherwise it is zero.

2.2. Weighted KNN imputation and SKNN imputation

In cluster-based estimation, MVs are estimated by combining the expression levels of *K*-nearest genes chosen based on a given similarity measure. Thus, KNN predictions are based on the intuitive assumption that objects close in distance are potentially similar. Both the measure to use for computing similarities between genes and the number of nearest neighbours (K) must be determined.

For a given target gene $\boldsymbol{x}_i$, KNNimpute (Troyanskaya et al., 2001) calculates a weighted Euclidean distance d_{ik} between the target gene i and each candidate gene k using the expression:

$$d_{ik} = \sqrt{\frac{\sum_{j=1}^{n} r_{ij} r_{kj} (x_{kj} - x_{ij})^2}{\sum_{j=1}^{n} r_{ij} r_{kj}}} \tag{1}$$

where r_{ij} is the element in the ith row and jth column of the missing indicator matrix $\boldsymbol{R}$. The missing entry j of target gene i is then estimated by the weighted average of the expression values of the K most similar genes in experiment j:

$$\hat{y}_{ij} = \sum_{k=1}^{K} w_{ik} x_{kj} \tag{2}$$