Missing Microarray Data Estimation Based on Projection onto Convex Sets Method

Author

Gan, XC, Liew, AWC, Yan, H

Published

2004

Conference Title

PROCEEDINGS OF THE 17TH INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION, VOL 3

DOI

10.1109/ICPR.2004.1334645

Missing Microarray Data Estimation Based on Projection onto Convex Sets Method

Xiangchao Gan[1], Alan Wee-Chung Liew[1] and Hong Yan[1,2]
[1]Department of Computer Engineering and Information Technology
City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong
[2]School of Electrical and Information Engineering
University of Sydney, NSW 2006, Australia

Abstract

DNA microarrays have gained widespread uses in biological studies. Missing values in a microarray experiment must be estimated before further analysis. In this paper, we propose a projection onto convex sets based algorithm to incorporate all a priori knowledge about missing values into the estimation process. Two convex sets applicable to all microarray datasets are constructed based on singular value decomposition (SVD). In addition, in the two most popular missing value estimation methods KNNimpute and SVDimpute, there is a trade-off whether to use a specific group of genes for the missing value estimation or to use all genes. Our algorithm can provide an optimal combination of these two strategies. Experiments show our algorithm can achieve a reduction of 16% to 20% error than the KNNimpute and SVDimpute methods.

1. Introduction

DNA microarray is a relatively new and rapidly advancing technology. Gene expressions measured using microarrays usually suffer from the missing value problem. Missing values occur due to various reasons, including artifacts on the microarray, insufficient solution and image corruption. Unreliable spots on a microarray image are usually manually flagged and excluded from subsequent analysis, resulting in missing data on those locations. However, in many analysis methods, the complete matrices are required. For example, the inability of many cluster algorithms to process the missing values necessitates the imputation of missing values.

Several methods have been suggested to deal with the missing value problem. Simple methods such as replacing missing values by zeros or by the averages of the expression profiles are often used. More advanced techniques, such as K-nearest neighbor method (KNNimpute) or the singular value decomposition method (SVDimpute), have recently been proposed in [1].

If an erroneous missing value replacement is performed, genes containing a high number of missing values can be classified incorrectly in subsequent cluster analysis. Thus, pattern recognition techniques for microarray data processing, such as k-means clustering and self-organizing maps, may benefit from using more accurately estimated missing values. By incorporating all information available about the missing values in estimation process, we can obtain an optimal result. The projection onto convex sets (POCS) algorithm provides such a convenient framework to solve this problem.

In this paper, the theory of POCS will be introduced first. Then, in Section 3, we analyze the SVD based method of [1], and propose two convex sets and the POCS estimation algorithm. In Section 4, we provide the experiment results of our algorithm and compare them with those obtained from existing methods. Finally, conclusions are given in Section 5.

2. Projection onto convex sets

In a POCS-based algorithm, every known *a priori* property about the original signal can be formulated as a corresponding convex set in a Hilbert space H [3]. Given n closed convex sets $C_i, i = 1,2,\cdots,m$, and nonempty $C_0 = \bigcap_{i=1}^{m} C_i$, then the iteration

$$\forall_n \in N \quad a_{n+1} = a_n + \lambda_n (P_{n(\mathrm{mod}\,ulo\ m)+1}(a_n) - a_n) \qquad (1)$$

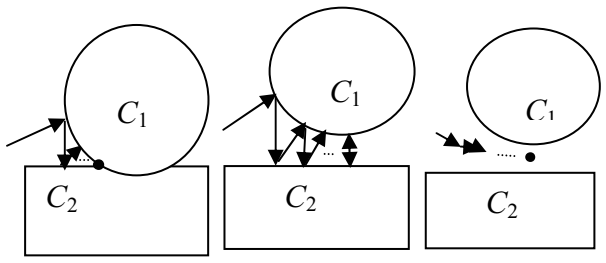
will converge to a point in set C_0 for any initial a_0, where P_i is the projector onto C_i defined by

$$\| x - P_i(x) \| = \min_{g \in C_i} \| x - g \| \qquad (2)$$

$\lambda_i \in (0,2)$ is the relaxed parameter and we often use $\lambda_i = 1$ for simplicity.

The method in Equation (1) is often called sequential projection. When the sets are non-intersecting, i.e., C_0 are empty, the iteration may fall into a trap. We illustrates it in Fig.1

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE



(a) (b) (c)

Fig. 1: The POCS iteration algorithm a) iteration in a sequential projection for a consistent problem, b) iteration in sequential projection for an inconsistent problem, c) iteration in simultaneous projections for a inconsistent problem.

In simultaneous projection algorithms, Equation (1) is substituted by

$$\forall_n \in N \quad a_{n+1} = a_n + \lambda_n \left(\sum_{l=1}^{m} w_l P_l(a_n) - a_n \right) \quad (3)$$

where w_l are the weights on the projections satisfying $\sum_{l=1}^{m} w_l = 1$ and $w_l > 0$ for all l. The simultaneous algorithm will always converge and provide a better solution for an inconsistent problem.

3. POCS-based imputation algorithm

In a microarray experiment, the relative expression levels of M genes of a model organism are probed simultaneously by a single microarray. A series of N arrays probe the genome-wide expression level in N different sample, i.e., under N different experimental conditions. We often use a matrix A of size $M \times N$ to tabulate the full expression data. If we perform singular value decomposition to matrix A, we get

$$A_{M \times N} = U_{M \times M} \Sigma_{M \times N} V^T_{N \times N} \quad (4)$$

Let $L = \min\{M, N\}$, matrix V^T now contains L eigengenes, and matrix U contains L eigenarrays [4]. In the SVDimpute of [1], the k most significant eigengenes from V^T are selected, and missing value j in gene i is estimated by first regressing this gene against the K eigengenes and then use the coefficients of the regression to reconstruct j from a linear combination of the K eigengenes.

Assume that w_i $i = 1,2,\cdots k$ are the eigengenes. For gene x, assume the value of its *j-th* component is missing. By deleting the *j-th* component of w_i and x, we get $\tilde{w}_i$ and $\tilde{x}$, then the missing component is estimated by

$$x(j) = \sum_{i=1}^{k} (\tilde{w}_i^t \tilde{x}) w_{i,j} \quad (5)$$

The POCS method requires the specification of convex sets that incorporate the a priori constraints about the solution. Based on the SVD of the expression data, we propose two convex sets for our POCS-based impute algorithm.

A. A Convex set using the eigengene information.

In SVD for genome-wide expression modeling, eigenvalue l given by $\Sigma(l,l)$ indicates the relative significance of the l-th eigengene and eigenarray in term of the fraction of the overall expression that they capture. In SVD-based imputation algorithm, only the K most significant eigengenes is used for the missing values estimation. That means we regard the other L-K eigenexpression as noise. The signal-to-noise ratio is given by

$$p = \sum_{k=1}^{K} \varepsilon_k^2 \Big/ \sum_{k=K+1}^{L} \varepsilon_k^2 \quad (6)$$

It is unavoidable that the missing values are also affected by noise. So, when a solution $\tilde{y}$ for a missing value is found using SVD-based method, a more reliable estimate is that the missing value lies in an interval $[(1 - w\sqrt{p})\tilde{y}, (1 + w\sqrt{p})\tilde{y}]$ where w is a parameter determined from the confidence limit.

Defining the positions in matrix A of all missing values as a set I, a convex set can be constructed as

$$C_v = \{B : \varepsilon_1 \tilde{A}(l) \le B(l) \le \varepsilon_2 \tilde{A}(l), l \in I\} \quad (7)$$

where the $\tilde{A}(l)$ is the estimation of a missing value using SVD method based on eigengenes, $\varepsilon_1 = (1 - w\sqrt{p})$ and $\varepsilon_2 = (1 + w\sqrt{p})$.

Given an initial value B, the projection onto set C_v can be given as

$$P_v(B(l)) = \begin{cases} \varepsilon_1 \tilde{A}(l) & for\ B(l) < \varepsilon_1 \tilde{A}(l) \\ \varepsilon_2 \tilde{A}(l) & for\ B(l) > \varepsilon_2 \tilde{A}(l) \\ B(l) & otherwise \end{cases} \quad (8)$$

B. A Convex set using the eigenarray information.

In [1], only the eigengenes were used for missing values estimation. Based on the matrix theory, when the *K* most significant eigengenes can capture the gene expression of all genes in matrix *A*, the *K* most significant eigenarrays can also capture the gene expression of all arrays in *A*. We have

$$C_u = \{B : \delta_1 \hat{A}(l) \le B(l) \le \delta_2 \hat{A}(l), l \in I\} \quad (9)$$

where the $\hat{A}(l)$ is the estimation of a missing value using SVD method based on eigenarray, $\delta_1 = (1 - w\sqrt{p})$ and $\delta_2 = (1 + w\sqrt{p})$. The projection P_u has the similar form as in Equation (8).

C. The POCS algorithm for missing value estimation

Using both the convex sets defined above, the POCS theory yields the following recovery algorithm:

1. Select a initial estimation x_0.
2. For $k = 1,2,\cdots,$ compute x_k from
   $$x_k = P_u P_v x_{k-1}$$
   where P_u, P_v denote the projectors onto the constraint sets C_u, C_v ,respectively.
3. If $x_k = x_{k-1}$, exit the iteration, else go to step 2.

In POCS algorithm, it is necessary to find a good initial point. According to the POCS theory, the algorithm always converges to a point in the intersection of all sets and with the least distance to the initial point. When a smoothest value is used as the initial point for missing value estimation, the solution of our algorithm is the smoothest one among all solutions which satisfy all *a priori* knowledge. To get the smoothest initial value for the time series dataset, the spline interpolation can be used; while for non-time series dataset, the average of the gene expression profile is a good choice.

In the KNNimpute algorithm of [1], only the closest *N* genes are used to estimate the missing values. This reduces the influence of irrelevant genes on the estimated missing values. In contrast, the SVD-based method is able to utilize all the gene profile correlation information while neglecting noise. The two methods manifest an apparent trade-off between local and global information and their combination becomes an attractive alternative. In our POCS method, we can combine the merit of both methods. For set C_v, when using eigenarrays to obtain the SVD-based missing value estimation, the complete matrix and the expectation maximization method introduced by [1] are used. For set C_u, when using eigengenes to obtain the SVD-based missing value estimation of gene T, we first find *N* other genes, whose expressions are most similar to T. The eigengenes are then calculated from the *N* genes and missing value estimation is performed subsequently. Thus, the POCS algorithm allows us to conveniently combine both global and local information to obtain a better solution.

4. Experiments

In this section, we apply our method to several gene expression datasets of yeast cell-cycle from Spellman et al. (http://cellcycle-www.stanford.edu) [6]. It contains expression profiles for 6178 genes under different experimental conditions, i.e., cdc15, and cdc28, alpha factor and elutriation experiments. To assess the performance of our missing value estimation algorithm, we also compare the normalized RMS error of our estimation algorithm with the KNNimpute and SVDimpute methods described in [1]. For the KNNimpute and the convex set C_v, we use the 20 closest genes. The parameter *w* that is associated with both C_v and C_u should ideally be determined statistically by confidence limit. However, in our experiment, we simply set it to *w*=1.

The normalized RMS (NRMS) error is calculated as follows

$$NRMS = \sqrt{\frac{\sum_{i=1}^{n} (\hat{B}(i) - B(i))^2}{\sum_{i=1}^{n} (\hat{B}(i))^2}} \quad (10)$$

where the *B* is the original gene expression matrix and $\hat{B}$ is the estimation obtained by missing values estimation algorithm.

Table 1 shows the results of three missing value estimation methods when 10% of the data is missing. Our algorithm can achieve 16-20% less error than the two existing methods.

	KNNimpute	SVDimpute	Our algorithm
alpha	0.2514	0.2507	0.2271
cdc15	0.2439	0.2509	0.2139
cdc28	0.2631	0.2713	0.2470
elu	0.2130	0.2219	0.1981

Table 1: Comparison of three methods in terms of normalized RMS error for the missing values.

5. Conclusion

In the microarray missing value estimation problem, for different experiments and different genes, we often have different *a priori* knowledge about the missing value. We can incorporate all available information about the missing values into the estimation process to obtain an optimal result.

In this paper, we have introduced a POCS-based algorithm for missing value estimation. Two convex sets that are applicable to all microarray dataset are introduced. They are constructed based on singular value decomposition. In addition, for the two most powerful missing value estimation methods KNNimpute and SVDimpute, there is a trade-off for whether to use a specific group of genes or to use all genes for missing value estimation. Our algorithm can provide a best combination of these two strategies. Experiments indicate that our algorithm is able to achieve between 16% to 20% reductions in estimation error when tested on the yeast cell-cycle datasets.

.

6. Acknowledgment

This work is supported by a CityU interdisciplinary grant (project 9010003).

7. References

[1]. O. Troyanskaya, M. Cantor G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman., "Missing values estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, pp.520-525, 2001

[2]. G. H. Golub and C. F. Van Loan, *Matrix Computations*, Baltimore, MD: Johns Hopkins Uinversity Press, 1996, 3rd ed.

[3]. H. Stark, and Y. Yang, "Vector Space Projections, A Numerical Approach to Signal and Image Processing, Neural Nets, and Optics", John Wiley& Sons, New York, 1998.

[4]. O. Alter, P. O. Brown and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling", *Proc. Natl Acad. Sci. USA*, 97, pp.10101-10106.

[5]. Z. B. Joseph, G. K. Gerber, D. K. Gifford, T. S. Jaakkola, and I. Simon, "Continuous Representations of Time-Series Gene Expression Data", *Journal of Computational Biology*, vol.10, pp. 341-356, 2003

[6]. P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher, "Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization", *Molecular Biology of the Cell*, Vol. 9, pp.3273-3297, December 1998