S5-5

Reuse of Imputed Data in Microarray Data Analysis

Gwan-Su Yi

Computational Biology Laboratory, Information and Communications University

The analysis of microarray data usually includes many statistical analyses and clustering algorithms which require complete data set. Microarray data include, however, missing values for diverse reasons of imperfection in the steps of microarray experiments such as the insufficient resolution, image corruption and so on. One of the yeast microarray data sets, which are used in this work, shows that the number of genes having at least one missing value is 2419 of 6198 rows (genes) (in other words, 39 %) and another one includes 2229 of 6718 genes (33.2%) containing at least one missing value. A few number of imputation methods for DNA microarray data have been introduced, but the efficiency of imputation of those methods was limited both in accuracy and computational complexity (Troyanskaya et al, 2001). The existence of missing values in a gene limits the use of other observed values of that gene in the conventional imputation method. In our work, this problem could be improved by using the imputed values sequentially for the later nearest neighbor calculation and imputation. We suggest a sequential KNN (SKNN) imputation method that boasts improved accuracy in estimation of missing values in a wide range of missing rates with high computational speed.

We developed a new cluster-based imputation method called sequential K-nearest neighbor (SKNN) method. This imputes the missing values sequentially from the gene having least missing values, and uses the imputed values for the later imputation. Although it uses the imputed values, the efficiency of this new method is greatly improved in its accuracy and computational complexity over the conventional KNN based method and other methods based on maximum likelihood estimation. The performance of SKNN was in particular higher than other imputation methods for the data with high missing rates and large number of experiments. Application of Expectation Maximization (EM) (Caruana R., 2001) to the SKNN method improved the accuracy, but increased computational time proportional to the number of iterations. The Multiple Imputation (MI) method, which is well known but not applied previously to microarray data, showed a similarly high accuracy as SKNN method, with slightly higher dependency on the types of data sets.

We evaluated the efficiency of our newly developed imputation method and the EM-SKNN method with three other imputation methods: KNN-based imputation, the MLE method, the MI method, by applying them to three different types of microarray data sets with different missing rates. The data sets were from a study of gene expression in yeast *Saccharomyces cerevisiae* cell-cycle regulation (Spellman et al.,1998), calcineurin/crz1p signaling pathway(Yoshimoto et al., 2002), and certain environmental changes(Gasch et al.,2000). These data sets can be classified as time series data set, mixed (time-series and non-time series) data set and non-time series data set. The efficiencies of imputation methods were assessed by the differences of Root Mean Squared (RMS) error and correlation coefficients using three different data types.

The accuracies of our new methods are especially superior when the missing rate is over 30% (Figure 1). The slight difference of KNN algorithm could lead to a large improvement in the accuracy of imputation at a high missing rate because the SKNN method is able to select more similar k-neighbors than the conventional KNN method as the missing rate grows. In the conventional KNN method, the selection pool and the dimension (or number of existing values for a gene) of the distance measurement of neighbor genes are reduced according to the increase of missing rate. In this situation, the method inevitably selects less related (or less similar) neighbors for imputation.

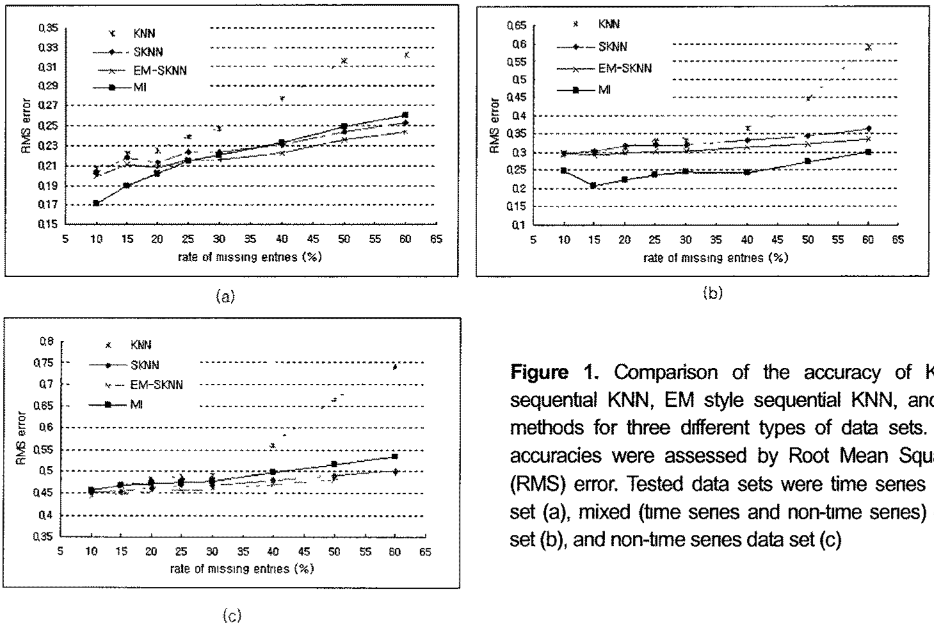

Figure 1. Comparison of the accuracy of KNN, sequential KNN, EM style sequential KNN, and MI methods for three different types of data sets. The accuracies were assessed by Root Mean Squared (RMS) error. Tested data sets were time series data set (a), mixed (time series and non-time series) data set (b), and non-time series data set (c)

We tested the performance of other well known non-KNN-based methods such as maximum likelihood estimation (MLE) and multiple imputations (MI) methods. These methods are well known imputation methods but there has been no report on their application in microarray data analysis. The efficiency of the MLE method was much worse than the SKNN method for all tested data sets. The efficiency of the MI method was better at a lower missing rate, but slightly worse at a higher missing rate for the time-series data set. MI method was as efficient as the SKNN method for the imputation of microarray data, but the efficiency of the MI method was more fluctuated than the SKNN method depending on the data type.

The computational complexity is reduced in the SKNN method for the dimension of both the number of genes and the experiments compared with the simple KNN method. Particularly, computation time can be saved substantially for microarray data with a large number of experiments. Although the SKNN method works efficiently in a wider range of missing rate with high speed, it is especially efficient on the data having high missing entries. It should be practically useful to save the data of some accidental microarray experiments having high missing entries.

We want to emphasize that our results showed that the method using estimated values achieved even better accuracy than the method using only observed values in the case of the KNN-based imputation method. It would be hardly acceptable for the experimentalist to use imputed data for further analysis. However, analysis could become more erroneous without imputation due to loss of information caused by missing values. The use of imputed data should definitely depend on the type of later process. If the next process is a cluster-based analysis, the genes with imputed values could be efficiently used, as we had good results for KNN-based imputation with the reuse of imputed values. Our results suggest that the imputed values generated by SKNN method can be used reliably for further cluster based analysis of microarray data.

References

1. Troyanskaya, O., Cantor, M., Sherlock, G.., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics.* 17, 520-525.
2. Caruana R., 2001. A Non-Parametric EM-Style Algorithm for Imputing Missing Values. *AI and STATISTICS 2001.*
3. Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. (2000) Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes, *Molecular Biology of the Cell*, Vol. 11, 4241-4257, December 2000.
4. Spellman, P. T., Sherlock, G., Zhqng, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998) Comprehensive Identification of Cell-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridzation. *Molecular biology of the Cell*, 9, 3273-3297
5. Yoshimoto, H., Saltsman, K., Gasch, A. P., Li, H. X., Ogawa, N., Botstein D., Brown, P. O., and Cyert, M. S. (2002) Genome-wide Analysis of Gene Expression Regurated by the Calcineurin/Crzlp Signaling Pathway in *Saccharomyces cerevisiae. The Journal of Biological Chemistry*, Vol. 277, No. 34, 31079-31088