Proceedings of Korean Data And Information Science Society,
October 28-29, 2005, pp. 153-159

Metastasis Related Gene Exploration Using TwoStep Clustering for Medulloblastoma Microarray Data

Sung-Su Ban[1], Hee-Chang Park[2]

Abstract

Microarray gene expression technology has applications that could refine diagnosis and therapeutic monitoring as well as improve disease prevention through risk assessment and early detection. Especially, microarray expression data can provide important information regarding specific genes related with metastasis through an appropriate analysis. Various methods for clustering analysis microarray data have been introduced so far. We used twostep clustering fot ascertain metastasis related gene through t-test. Through t-test between two groups for two publicly available medulloblastoma microarray data sets, we intended to find significant gene for metastasis. The paper describes the process in detail showing how the process is applied to clustering analysis and t-test for microarray datasets and how the metastasis-associated genes are explorated.

keywords : microarray, twostep clustring, medulloblastoma, metastasis

1. Introduction

DNA microarray technology allows simultaneous monitoring of the expression levels of numerous genes. There are numerous potential clinical application that could refine diagnosis and therapeutic monitoring as well as improve disease prevention through risk assessment and early detection. This technology is being more and more widely applied in biological and medical research to address a wide range of questions.

Microarray experiments generate large and complex multivariate data sets, and some of the greatest challenges lie not in generating these data but in the development of computational and statistics tools to analyse the large

[1] Graduate Student, Department of Bioinformatics, Changwon National University, Changwon, Gyungnam, 641-773, Korea
[2] (Corresponding author) Professor, Department of Statistics, Changwon National University, Changwon, Gyungnam, 641-773, Korea
E-mail : hcpark@sarim.changwon.ac.kr

amounts of data(Yang and Speed, 2002). The common goal is to find groups of genes that are coordinately or differentially expressed, and to find groups of samples that are biologically relevant. A number of different methods have been used for clustering, but the twostep cluster methos is a scalable cluster analysis algorithm designed to handle very large datasets. We performed twostep clustering algorithms included in Clementine with the two medulloblastoma microarray data sets publicly available on online.

The remainder of this article is organized as follows: In Section 2, we explain the general knowledge about microarrays. In Section 3, we give an brief review of TwoStep clustering analysis. In Section 4, we describe data construction for variables selection and twostep clustering on the basis of this variables. In Section 5, we repesent the process and result.. We conclude in Section 6.

2. DNA Microarray

A DNA microarray consists of a solid surface, usually a microscopic slide, onto which DNA molecules have been chemically bonded. The purpose of a microarray is to detect the presence and abundance of labelled nucleic acids in a biologic sample, which will hybridise to the DNA on the array via Watson-Crick duplex formation, and which can be detected via the label(Stekel, 2003).

In the majority of microarray experiments, the labelled nucleic acids are derived from the mRNA of a sample or tissue, and so the microarray measures gene expression. Especially, it is called complementary DNA(cDNA) that signifies single-stranded labelled DNA synthesized in the laboratory using messenger RNA as a template. Microarray technology is extensively used in biological research. There are currently two main trends in microarray technology, cDNA bicolar glass slides(Chee et al., 1996) and the high-density oligonucleotide arrays manufactured by Affymetrix(Lipshutz et al., 1999). The applied technologies vary greatly between laboratories. The power of a microarray is that there may be many thousands of different cDNA molecules bonded to an array, and so it is possible to measure the expression of many thousands of genes simultaneously.

In conclusion, DNA microarray is high-throughput gene expression analyis system through searching cDNA derived from the mRNA of a biologic sample. Monitoring gene expression lies at the heart of a wide variety of medical and biological research projects, including classifying diseases, understanding basic biological processes, and identifying new drug targets. Until recently, comparing expression levels across different tissues or cells was limited to tracking one or a few genes at a time. Using microarray, it is possible to simultaneously monitor the activities of thousands of genes.

3. The TwoStep Clustering Algorithm

The SPSS TwoStep cluster can handle both continuous and categorical variables or attributes. It has two steps 1) precluster the cases (or records) into many small sub-clusters; 2) cluster the sub-clusters resulting from pre-cluster step into the desired number of cluster. Two distance measures, log-likelihood distance measure and euclidean distance are available in both steps.

3.1 Step 1: Pre-cluster

The pre cluster procedure is implemented by constructing a modified cluster feature (CF) tree. The CF tree consists of levels of nodes, and each node contains a number of entries. A leaf entry (an entry in the leaf node) represents a final sub-cluster. The non-leaf nodes and their entries are used to guide a new record quickly into a correct leaf node. Each entry is characterized by its CF that consists of entry's number of records, mean and variance of each continuous variable, and counts for each category of each categorical variables. Upon reaching a leaf node, it finds the closest leaf entry in the leaf node. If the record is within a threshold distance of the closest leaf entry, it is absorbed into the leaf entry and and the CF of that leaf entry is updated. Otherwise it starts its own leaf entry in the leaf node. If the CF-tree grows beyond allowed maximum size, the CF-tree is rebuilt CF-tree is smaller and hence has space for new input records. This process continues until a complete data pass is finished.

3.2 Step 2: Cluster

The cluster step takes sub-clusters resulting from the pre-cluster step as input and then groups them into the desired number of cluster. The number of clusters depends on the data set at hand. In the first step, the BIC(Bayesian information criterion) or AIC(Akaike information criterion) for each number of clusters within a specified range is calculated and used to find the initial estimate for the number of clusters. In the second step, the initial estimate is refined by finding the largest increase in distance between the two closest clusters in each hierarchical stage. The BIC and AIC for J clusters are defined as

$$BIC(j) = -2\sum_{j=1}^{j}\xi_j + m_j \log(N) \tag{3.1}$$

$$AIC(j) = -2 \sum_{j=1}^{j} \xi_j + 2m_j \ . \tag{3.2}$$

Where

$$m_j = J\{2K^A + \sum_{k=1}^{K^B}(L_k - 1)\},$$

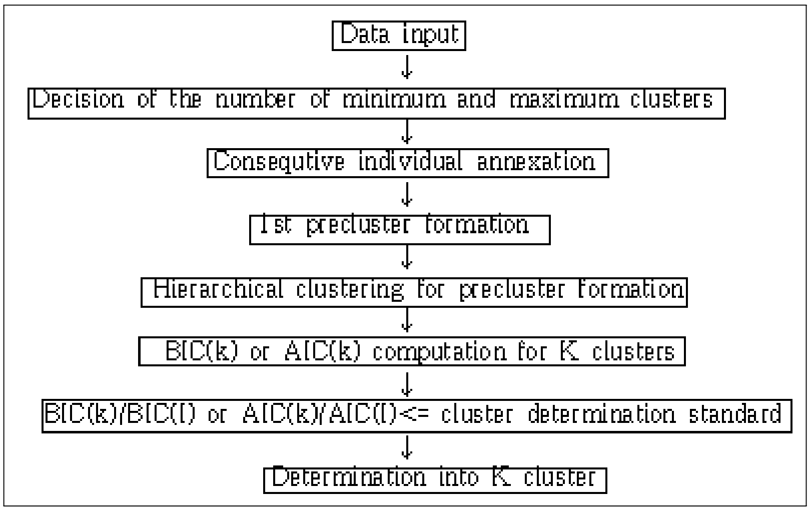
(3.3)

K^A : Total number of continuous variables used in the procedure,

K^B : Total number of categorical variables used in the procedure,

L_k : Number of data records in cluster k.

The likelihood function is utilized assuming normal distribution for continuous variables and multinomial distributions for categorical variables in twostep. The twostep clustering algorithm is as the following figure 1.


Data input
Decision of the number of minimum and maximum clusters
Consequtive individual annexation
1st precluster formation
Hierarchical clustering for precluster formation
BIC(k) or AIC(k) computation for K clusters
BIC(k)/BIC(l) or AIC(k)/AIC(l)<= cluster determination standard
Determination into K cluster

<Figure 1> TwoStep clustring Algorithm

4. Data Construction

4.1 Medulloblastoma

Medulloblastoma, a highly invasive primitive neuroectodermal tumor of the cerebellum, is the most common malignant brain tumor of childhood. Approximately one-third of patients presents with metastatic disease at the time of diagnosis, and up to two-thirds will have leptomenigeal disease at the time of relapse. Current management strategies combine surgery, radiotherapy, and chemotherapy, and long-term survival rates are 60% to 80%. Unfortunately, cognitive deficits and other sequelae of therapy are common among survivals(Packer, 1999). Thus, understanding the genetic regulation of the signaling pathways may greatly impact on the development of more effective and less toxic therapies.

4.2 The Medulloblastoma Affymetrix GeneChip Data

To investigate metastasis-associated gene, we used two publicly available medulloblastoma microarray data sets. The total of 83 medulloblastoma specimens were expression profiled by Pomery et al.(2002)(Affymetrix HuGenFl array; approximately 6,800 gene transcripts) and MacDonald et al.(2001)(Affymetrix G110 array; approximately 2,000 gene transcripts). The raw data of this microarray experiments are available online (http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi) and (http://microarray.cnmcresearch.org) respectively. The combined data stes are comprised of 28 metastatic (M+)(18 specimens by Pomery et al and 10 specimens expression profiled by MacDonald et al respectively) and 55 nonmetastatic (M0) tumors(42 and 13 expression profiled respectively).

.

5. Analysis

5.1 Analysis process

The investigation process for metastasis related gene exploration using twostep clustering is like the following figure 2. In this article, we utilized the Clementine 8.0 of SPSS for construction of model to address this question.

[Process 1] Variable selection

The variable for clustering the other data sets is selected through t-test for each gene of two groups(10 metastatic (M+) data stes and 13 nonmetastatic (M0) data sets expression profiled by MacDonald et al.

[Process 2] Decision of the cluster numbers

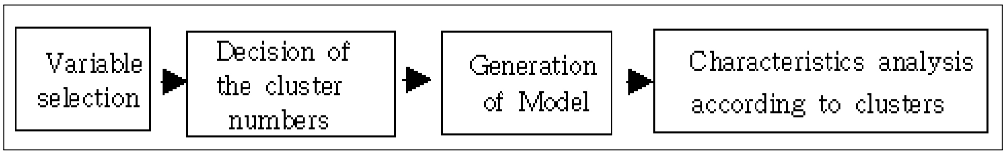
Twostep clustering algorithm finds the optimal number of clusters within the user defined minimum and maximum limits. We defined the maximum limit as 14 and the minimum limit as 2.

[Process 3] Generation of Model

Twostep clustering analysis model is generated on the basis of selected variables. With the result, two clusters were generated.

[Process 4] Characteristics analysis according to clusters

Characteristics is analyzed by clusters. The characteristics of two clusters was consistent with the characteristics classified in anticipation.

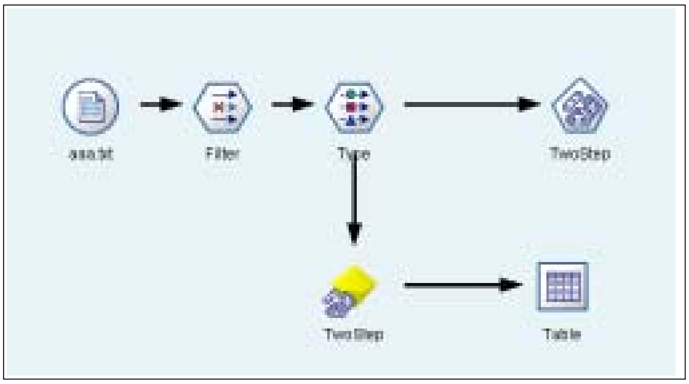

Variable selection
Decision of the cluster numbers
Generation of Model
Characteristics analysis according to clusters

<Figure 2> Model construction and analysis process for metastasis

6 Result

5.1 Twostep cluster analysis stream

The generating stream of twostep clustering is like figure 3.



<Figure 3> The generating stream of twostep clustering

5.2 Metastasis related gene seletion through t-test

First of all, we selected some significant genes through t-test in significan level less than 0.01 using two microarray data set. The eight genes were selected in microarray data by MacDonald et al.(2001) and the fifty three genes were selected as the same level in microarray data by Pomery et al.(2002).

5.3 Twostep clustering on the basis of significant genes

Two microarray data sets were not clustered properly using cotinuous data of those microarray data on the basis of selected gene in significant leve. That is, the patients group had metastatic disease and the patients group did not have metastatic disease were discriminated in the appropriate manner when we had used continuous data. Therefor, we clustered the binary transformed data of the same microarray using twostep. With the result the better clustered result was produced.

6. Conclusion

In this article, we applied twostep clustering to the medulloblastoma DNA microarray data to discern metastasis related gene. We found somr cell proliferating gene and oncogene. In using microarray data, binary tranformed data would produce better results. Becuase the microarray data has many error itself. But, the results of this study need further study in wet laboratory and conformation with experimental results. Especially, due to the small number of microarray data, there is a limit to accept the generated facts as general rule. With the replenishment of more data and the introduction of various kind of mining technique, still better result could be found.

References

1. MacDonald, T.J., Brown, L., LaFleur, B., Petrson, K., Lawlor, C., Chen, Y., Packer, R.J., Cogen, P., Stephan, D.A.(2001). Expression profilling of medullobastoma: PDGFRA and the RAS/MAPK pathway as therapeutic targets for metastatic disease. Nat Genet, Vol 29, pp. 143-152.
2. Nielsen, G.G., Gill, R. D., Andersen, P. K., and Sϕrensen, T. I. A. (1992). A Counting Process Approach To Maximum Likelihood Estimation In Frailty Models. *Scandinavian Journal Of Statistics*, Vol 19, pp. 25-43.
3. Pomery, S. L. et al. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*. Vol 415, pp. 436-442.
4. Yang, Y. H. and Speed, T. (2002). Design Issues for cDNA Microarray Experiments. *Nature Genetics.* Vol 3, pp. 579-588.
5. Zhang, T., Ramakrishnon, R., Livny M. (1996). BIRCH: An Efficient Data Clustering method for Very Large Databases. *Proceedings of the ACM SIGMOD Conference on Management of Data,* pp. 103-114.
6. The SPSS TwoStep Cluster Component.www.spss.com.