http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (23): 7898-7901, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.7898.7901

Bicluster Significance Evaluation with the Application of Information Entropy

[1]Zhang Yanjie, [2]Hu Zhanyi and [1]Sun Limin
[1]School of Computer Science and Technology, Yantai University, 264005, Shandong, China
[2]Institute of Automation, Chinese Academy of Sciences, 100080, Beijing, China

Abstract: Along with the development of the technology of microarray chips, more and more gene expression data are available. Biclustering with gene expression data has been proved to be an efficient way to discover the characteristic genes corresponding to some specific diseases. It also has wide applications in the other areas. Since there are usually quite a lot of different size biclusters lying in the original data matrix and not all of the biclusters play the same roles, how to evaluate the significance of the detected biclusters is very important. Information Entropy (IE) as a way to measure the uncertainty in a random variable is vital in information theory. In this study we propose a method of applying self-defined IE as an index to evaluate the significance of all the detected biclusters, based on it the significance of each bicluster can be quantified. The number of useful biclusters can be greatly decreased while keeping the high recognition accuracy. The preliminary experiment results shown at the end of the study demonstrate its feasibility.

Key words: Microarray, gene expression data, biclustering, rule extraction

INTRODUCTION

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. These products are often proteins, but in non-protein coding genes such as ribosomal RNA (rRNA), transfer RNA (tRNA) or small nuclear RNA (snRNA) genes, the product is a functional RNA. Since an array can contain tens of thousands of probes, a microarray experiment can accomplish many genetic tests in parallel. Gene expression data is a kind of data matrix used to represent the expression level of many different genes under specific conditions simultaneously.

Usually in the data matrix, genes are arranged in the row direction while the column direction represents different time or different environmental conditions. Each element is a real number which is often the logarithm of the relative abundance of the mRNA of the gene (Madeira and Oliveira, 2004). Determining if changes in gene expression are statistically significant between different conditions, e.g., two different tumor types and determining the biological function of the differentially expressed genes, are important aims in a microarray experiment.

A bicluster is a submatrix of a data matrix (Hartigan, 1972). The difference between a bicluster and a submatrix is all the biclusters are definitely submatrices, for a data matrix whose size is m×n, there are totally (2^m-1) (2^n-1) ways to form its submatrices but only those submatrics whose row or column vectors satisfying some kind of linear relations will be treated as biclusters. When the data matrix is given, the number of biclusters lying in the data varies along with the data structure. Furthermore the sizes and the spatial positioning relations among those biclusters are completely unknown and strongly data dependent.

Over the last decade, biclustering methods have become more and more popular in different fields of two way data analysis and a wide variety of algorithms and analysis methods have been published (Prelic *et al*., 2006). The authors (Kaiser and Leisch, 2008) introduced the R package which contains a collection of bicluster algorithms, preprocessing methods for two way data and validation and visualization techniques for bicluster results. Biclustering algorithms can also be applied to classi?cation problems or decision support systems in various application ?elds, such as fault detection, biology and medicine (Alon *et al*., 1999; Alizadeh *et al*., 2000; Pomeroy *et al*., 2002). The detected biclusters provides a way of better interpretability of the results and provides more insight into the classifier structure and decision making process (Roubos *et al*., 2003).

BICLUSTER SIGNIFICANCE EVALUATION

Usually the number of biclusters lying in the original data matrix is huge when all these biclusters are detected with some kind of biclustering method, how to select the significant biclusters is very important. In the real situation, when the experiment data is given, most of the

Corresponding Author: Zhang Yanjie, School of Computer Science and Technology, Yantai University, 264005, Shandong, China

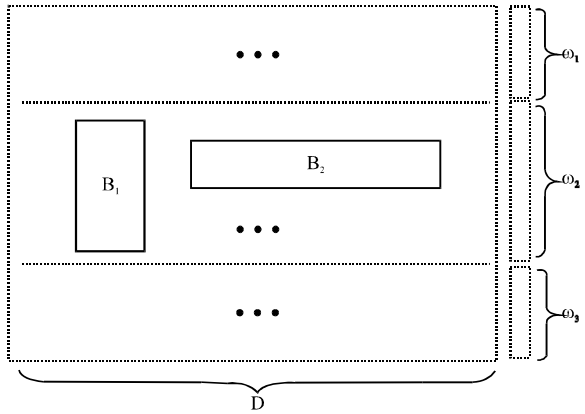

Fig. 1: Biclustering results of the original data matrix D with the class labels added as the last column

case the relation between the samples and the classes are known. Directly biclustering with the original data matrix D doesn't help to find the biclusters existing only in the samples of the same class. Whatever these biclusters are very helpful to disclose the interval relations among the same class so as providing information to do samples classifying.

Just as shown in Fig. 1, the class labels ω_1, ω_2, ω_3 corresponding to the samples are added as the last column of D. The newly generated data matrix which is factually the matrix column extension result with D is called E (D). Biclustering with E (D) will output a great amount of biclusters. Here only the bicluster B_1 and B_2 are used for analysis without considering the other biclusters. These two biclusters are within the samples of class ω_2. The difference between B_1 and B_2 is that B_1 covers more row numbers while B_2 covers more column numbers. For a recognition problem, B_1 gives the information of using few attributes to identify the samples which should be classified into class ω_2. As to the bicluster B_2, it just tells that the elements within itself are quite similar with each other. So the significance of B_1 is more important than that of B_2. Undoubtedly when the relation between the samples and the classes are known, biclusters such as B_1 will be more useful to be treated as an characteristic to indicate which class the newly input sample should be classified into. However, when the number of detected biclusters is huge, how to evaluate the significance of founded biclusters is very important.

Though most biclustering algorithms deal with the original numerical data matrix directly, in fact fuzzification can be done with the original data matrix before they enter future processing. During this procedure, each column of the data matrix can be considered as a fuzzy variable. If the size of the data matrix is m×n, then there are totally mn membership of n fuzzy variable needed to be confirmed. We can also give a reasonable explanation for this kind of processing. In the real life the exploration of rules can't always be built on the data that meet the exact mathematical relations. So this kind of data fuzzification will make it be easy for further work. And the final result will be more appropriate to the human being's thinking style. It is also very important in data mining.

Beside the data fuzzification method, data discretization is another kind of technique which is used to partition continuous attributes into a finite set of adjacent intervals in order to generate attributes with a small number of distinct values. Assuming that a dataset consisting of M examples and S target classes, a discretization algorithm would discretize the continuous attribute A in this dataset into a serial n discrete intervals $\{[d_0, d_1], (d_1, d_2], \cdots, (d_{n-1}, d_n]\}$, where, d_0 is the minimal value and d_0 is the maximal value of attribute A. Such a discrete result is called a discretization scheme on attribute A. This discretization scheme should keep the high interdependency between the discrete attribute and the target class to carefully avoid changing the distribution of the original data (Tsai *et al*., 2008).

If a bicluster B is composed of the row numbers' set $\{i_1, i_2, ..., i_m\}$ and column numbers' set $\{j_1, j_2, ..., j_n\}$ of the original data matrix D, then the function Φ (B) = $\{i_1, i_2, ..., i_m\}$ is defined to determine the set of those row numbers that the elements of B lie in D and $|\Phi(B)| = m$, so is the definition of the function Ψ (B) = $\{j_1, j_2, ..., j_n\}$.

Given a data matrix D, each row vector of D can be considered as a sample. Assume the number of samples lying in D is N, all these samples belong to different classes named ω_i, i = 1, 2, …, M. Define N_i is the number of samples in class ω_i, then p_i which means the probability of one sample belonging to the class ω_i can be estimated by N_i/N. The expected information entropy provided by D is:

$$I(N_1, N_2, \cdots, N_M) = -\sum_{i=1}^{M} p_i \log_2 p_i\}$$

If an attribute A has a number of k values which are $\{a_1, a_2, \ldots, a_k\}$, then the whole samples set can be classified into k different subset $S_1, S_2, \ldots, S_k$ by only using attribute A. Assume N_{ij} is the number of samples in subset S_j while these samples belonging to the class ω_i, then the information entropy of classified result by attribute A is defined as:

$$I(A) = \sum_{j=1}^{k}\left[\left(\frac{N_{1j}+N_{2j}+\cdots+N_{Mj}}{N}\right)I(N_{1j}, N_{2j}, \cdots, N_{Mj})\right]$$

Specially:

$$I(N_{1j}, N_{2j}, \cdots, N_{Mj}) = -\sum_{i=1}^{M} p_{ij} \log_2 p_{ij}$$

where, $p_{ij} = N_{ij}/|S_j|$ means the probability of samples lying in subset S_j belonging to the class ω_i. The whole information gain acquired by attribute A is:

$$Gain(A) = I(N_1, N_2, \cdots, N_M) - I(A)$$

Assume that there is a bicluster B when doing biclustering with E (D), as we know the more row numbers B covers and the less attributes B has, B is more meritorious, so we define:

$$\Delta(B) = \frac{|\Phi(\omega)|}{|\Phi(B)|}$$

as a weight to indicate the importance of information provided by B. $|\Phi(\omega)|$ is the number of samples belonging to the class ω where, the bicluster B is founded. The formula I (A) only instructs how to calculate the information entropy with one attribute. Since for any bicluster B it satisfies $|\Psi(B)| \geq 2$, that means we have to take a number of $|\Psi(B)|$ attributes' information entropy into account simultaneously.

For any two different attributes A_1 and A_2, if $I(A_1) < I(A_2)$, then the values taken by A_1 are more regular than the values taken by A_2. The values of attributes A_1 of the samples in the same class will be more stable than that of A_2. When applying these two attributes to classify an unknown input sample, the classified result based on A_1 will be more accurate than that of A_2. So for a bicluster B_1 whose column numbers' set is $\{A_1, \omega\}$ and another bicluster B_2 whose column numbers' set is $\{A_1, A_2, \omega\}$ while they have the same row numbers' set, the significance of B_1 should be bigger than that of B_2. That is to say, the information entropy of B_1 is smaller than that of B_2 if we define the IE of the bicluster B as:

$$\max_{A \in |\Psi(B)|} Gain(A)$$

Considering all of the analysis mentioned above, we define the following formula as an index to evaluate the significance of the bicluster B:

$$IE(B) = \Delta(B) \max_{A \in |\Psi(B)|} Gain(A)$$

COMPUTATIONAL EXAMPLE

In order to illustrate the feasibility and effectiveness of the algorithm, we apply the well-known wine data as experiment data (De Oliveira, 1999). The wine data contains the chemical analysis of 178 wines produced in the same region in Italy but derived from three different cultivars. The problem is to distinguish the three different types based on 13 continuous attributes (Table 1) derived from chemical analysis: alcohol, malic acid, ash, alcalinity of ash, magnesium, total phenols, flavanoids, nonflavanoids phenols, proanthocyaninsm color intensity, hue, OD280/OD315 of diluted wines and proline.

Compared with the real gene expression data, the wine data has smaller dimensionality while they are all numerical data. The application of wine data will help to save a lot of time to verify the feasibility of the proposed algorithm without destroying the nature of the research data object.

Here the data discretization method proposed in (Tsai *et al*., 2008) is applied and the discretization schemes are listed in Table 1. Biclustering with the discretized wine data E (D), a number of 5716 biclusters are founded. Among all these biclusters, 217 biclusters are within ω_1, 3794 biclusters are within ω_2 and 1705 biclusters are within ω_3. The formula IE (B) is used to calculate the IE of each bicluster. Then the significance of all of the biclusters within the same class can be sorted in an ascending order. Since the bicluster with the smallest IE is more valuable, three biclusters belonging to the class ω_1, ω_2 and ω_3 individually are selected. The column numbers' sets of these three biclusters are $\Psi(B_1) = \{1, 3, 5, 11, 12, \omega_1\}$, $\Psi(B_2) = \{13, \omega_2\}$ and $\Psi(B_3) = \{3, \omega_3\}$. And the Ies of these three biclusters are IE (B_1) = 1.5668, IE (B_2) = 1.6603 and IE (B_3) = 1.5668.

For the three selected biclusters, B_1 and B_3 cover all of the row numbers within their class ω_1 and ω_3, respectively while B_2 covers 67 row numbers out of 71 the total row numbers within the class ω_2, so these three

Table 1: Discretization schemes of all the attributes of the wine data

Atrribute	d_0	d_1	d_2	d_3
Alcohol	11.03	12.780	14.830	
Malic acid	0.74	1.475	2.235	5.8
Ash	1.36	2.030	3.230	
Alcalinity of ash	10.60	17.900	30.000	
Magnesium	70.00	88.500	162.000	
Total phenols	0.98	1.840	2.335	3.88
Flavanoids	0.34	1.400	2.310	5.08
Nonflavanoid phenols	0.13	0.395	0.660	
Proanthocyanins	0.41	1.305	1.655	3.58
Color intensity	1.28	3.820	7.550	13.00
Hue	0.48	0.785	1.710	
OD280/OD315	1.27	2.115	4.000	
Proline	278.00	755.000	1680.000	

biclusters offer a recognition accuracy of 97.75% A percentage of 99.95% biclusters can be removed while keeping high recognition accuracy.

CONCLUSITION

In this study we propose a method of applying information entropy as a way to evaluate the significance of the great number of biclusters detected in the original data matrix. The significance of each bicluster can be quantified by its IE and sorted in an ascending order based on it. Those biclusters with small IEs and covering as many as row numbers are selected. By this way the number of useful biclusters can be greatly decreased while keeping the high recognition accuracy. Processing with real gene expression data is ongoing and will be presented in the future work.

ACKNOWLEDGMENTS

This study was supported by Shandong Province Young Scientists Award Fund No.BS2012SW023 and Shandong Province Young Backbone University Teachers Civil Visiting Scholars Project.

REFERENCES

Alizadeh, A.A., M.B. Eisen, R.E. Davis, C. Ma and I.S. Lossos *et al*., 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403: 503-511.

Alon, U., N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack and A.J. Levine, 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci., 96: 6745-6750.

De Oliveira, J.V., 1999. Semantic constraints for membership function optimization. IEEE Trans. Syst. Man Cybernetics A, 29: 128-138.

Hartigan, J.A., 1972. Direct clustering of a data matrix. J. Am. Stat. Assoc. 67: 123-129.

Kaiser, S. and F. Leisch, 2008. A toolbox for bicluster analysis in R. Technical Report No.28, Department of Statistics. http://epub.ub.uni-muenchen.de/3293/.

Madeira, S.C. and A.L. Oliveira, 2004. Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Trans. Comput. Biol. Bioinform., 1: 24-45.

Pomeroy, S.L., P. Tamayo, M. Gaasenbeek, L.M. Sturla and M. Angelo *et al*., 2002. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature, 415: 436-442.

Prelic, A., S. Bleuler, P. Zimmermann, A. Wille and P. Buhlmann *et al*., 2006. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics, 22: 1122-1129.

Roubos, J.A., M. Setnes and J. Abonyi, 2003. Learning fuzzy classification rules from labeled data. Inform. Sci., 150: 77-93.

Tsai, C.J., C.I. Lee and W.P. Yang, 2008. A discretization algorithm based on class-attribute contingency coefficient. Inform. Sci., 178: 714-731.