An Effective Soft Clustering Approach to Mining Gene Expressions from Multi-Source Databases

CHIEN-I LEE, HSIU-MIN CHUANG
Department of Information and Learning Technology
National University of Tainan
33, Sec. 2, Shu-Lin St., Tainan City 700
TAIWAN

Abstract: - In recent years, many technologies that are used to analyze genes were proposed. Huge amount of biological databases, such as microarray data, biomedical literatures, sequence data and genome structure data et al., have formed useful data warehouses to mine gene-gene relations and predict the gene networks in advance. In the field of bioinformatics, the clustering of gene expressions is a common technology to extract the new knowledge. However, to raise the accuracy of gene clusters is a challenge because of the errors of biological databases and divergence of various clustering methods. In this paper, Multi-Source Soft Clustering (MSSC), which is an integrated framework of the clustering methods and multi-source databases, is presented to raise the accuracy. Two soft clustering methods, fuzzy c-means and soft CAST, are applied to solve the questions that genes may have multi-functions and involve several biological pathways. Combining microarray data and biomedical literatures to improve the overall accuracy may be better than using only one single dataset. In addition, the MSSC adopts the concept of clustering before integrating, and uses the correlation coefficient in statistics to calculate the distances of the matrices between the diverse soft clustering results. The experimental result shows that MSSC approach can be relatively more effective.

Key-Words: - gene expressions, soft clustering, text mining

1 Introduction

With the growth of the biological technology, enormous biological databases are produced and formed a hot issue of bioinformatics. It creates a need and challenge for data mining. Data mining is a process of the knowledge discovery in databases and the goal is to find out the hidden and interesting information [8]. The technology includes association rules, classification, clustering, and evolution analysis etc. Clustering algorithms are used as the essential tools to group analogous patterns and separate outliers according to its principles that elements in the same cluster are more homogenous while elements in the different ones are more dissimilar [2]. Furthermore, clustering algorithms do not need to rely on the pre-defined classes and the training examples while classifying the classes and can produce the good quality of clustering, so they fit to analyze the unknown genes better. For the reason, clustering algorithms often are applied to analyze the experimental results and identify the important genes from the numerous dataset. For example, the cDNA microarray technology which monitors the expressions levels for ten of thousands of genes in parallel by the ratio of signal intensity between the test sample and the control sample [6]. Microarray data may include the unknown genes and the known genes in various experimental conditions, and clustering algorithms can find out the co-regulated genes, distinguish the normal and abnormal tissues. It helps us to understand the gene regulation, gene function and cellular process for the genome.

Besides microarray data, biomedical literatures are also abundant resources for mining the meaning information, such as MEDLINE and BIOSIS etc. Because it is difficulty to collect the important information directly from the unstructured and semi-structured documents, text mining is utilized to extract keywords, analyze the syntax to induce the rules and interpret the biological mechanism. From many published researches of document mining, text analysis of biomedical literature has also been applied successfully to incorporate function information about the analysis of genes and infer the rules. Consequently, combining the microarray data and biomedical literatures, even the other databases, such as sequence data and genome structure data, to mine more interesting rules is formed an important research issue in the field of the bioinformatics.

In the paper, an integrated framework of clustering methods and heterogeneous data, Multi-Source Soft clustering (MSSC) is proposed. The method mainly is suitable in the condition that genes may have

multi-functions, involve several biological pathways and present some characteristic of clusters. It utilizes the correlation coefficient in statistics to calculate the distances of different soft clustering results. Fuzzy exponents and the weight parameter are used to adjust the softness of clustering and integrate various databases according to the property of actual datasets.

The first idea to link microarray data and literatures was given by Masys [11]. Following the analogous concept, a hierarchical clustering method was used to cluster the microarray data and the extracted-terms from the literatures were strengthened to recognize the boundaries of the clusters [12]. Hu et al. [9] developed a new framework of the cluster ensemble to integrate different clustering algorithms and applied the text summarization to infer the meaningful relations. However, these approaches ignore the mutual relations of multiple databases. Glenisson et al. described an overview of combining gene expression data and literature information [7]. Three ways to integrate multiple sources were presented and the differences of ways depended on the early or late of clustering and integrating. The great challenge lies in integrating various data sources deeply into a learning framework rather than using or linking them dependently. Recently published studies have shown an algorithm of semantic integration of multiple sources, Multi-Source Clustering (MSC). MSC adopts a variant of EM method to stochastically build the model multi-sources [13, 14]. It combined the heterogeneous sources before trained cluster models. In the experiment, the clustering quality of MSC performed better than the meta-clustering method. However, most clustering algorithms belong to hard clustering. The question is ignored that the genes could be different actors in different conditions and may belong to multiple clusters. For the reason, it adopts the soft clustering algorithms could tend to lead to better clustering results for overlapping clusters.

2 Methods

The mutations of our MSSC approach are based on the concept of MSC approach. MSC is a generalized version of the standard K-means and stochastic exploration of cluster models from multiple sources by boosting the model with the cluster assignments. The hard clustering strategy is that the cluster assignment of each point for each source is thought as k-dimensional vector in which only one entry is equal to 1 and the others are zero. In the other word, each point is only assigned to a cluster. Therefore, soft clustering methods are to consider the question that genes can be assigned to multi-clusters to enable to conform to the actual behavior of genes. To raise the overall quality of clustering, there are three questions from previous methods as follows: (1) Hard clustering approaches may make reduce the accuracy of clusters while genes belong to multi-clusters. (2) Combining heterogeneous data before clustering could decrease the meaningful information while little amount of genes are important in the certain database. (3) Weight of integrating heterogeneous data sources should be adjusted-parameters rather than the same ratio. The purpose of MSSC is to raise the overall accuracy of gene clusters by two soft clustering algorithms, fuzzy c-means (FCM) and soft CAST. Fig. 1 represents an overview of the MSSC.

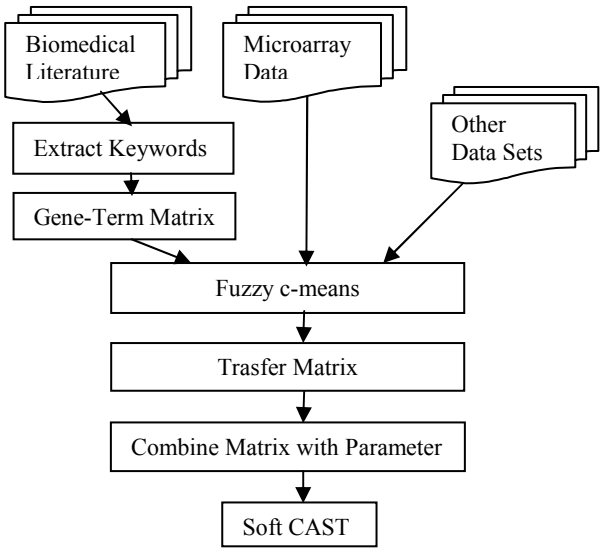

Fig.1 the overview of MSSC

2.1 Soft Clustering

A variety of clustering algorithms can be roughly classified five categories according to the function of clustering: hierarchical clustering, partitioning clustering, density-based clustering, grid-based clustering, and model-based clustering. Furthermore, these categories are divided into hard and soft clustering. Soft clustering algorithms are effectively applied in classifying groups including both the overlapping and non-overlapping clusters. Further, the fuzzy c-means algorithm is one of the most widely used soft clustering algorithms. It is a variant of standard k-means algorithm that uses a soft membership function. Given a set X of objects x_1, x_2,…, x_n, the fuzzy c-means algorithm tries to optimize a least-squared error function:

$$J_m(U, V) = \sum_{k=1}^{n}\sum_{i=1}^{c}(\mu_{ik})^m \| x_k - v_i \|^2$$

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{\| x_k - v_i \|}{\| x_k - v_j \|}\right)^{\frac{2}{m-1}}}$$

where u_{ik} is the membership x_k in the kth fuzzy cluster satisfying $\sum_{i=1}^{c} u_{ik} = 1, \forall k = 1,..., n$, and v_i is the cluster centroid as follows:

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^m x_k}{\sum_{k=1}^{n} u_{ik}^m}$$

The fuzzy factor m controls the fuzzy intensity of clustering results. As m approaches one, the clustering results are non-overlapping as well as the standard k-means results.

In the paper, the other soft clustering algorithm is soft CAST. Soft CAST is a variant of Cluster Affinity Search Technique (CAST) which was developed by Ben-Dor et al. to cluster gene expression data [1]. The CAST takes an input as a parameter called the affinity threshold t, where $0 < t < 1$, and tries to guarantee the average similarity in each generated cluster is higher than the threshold t. The clusters are constructed one, C_{open}, at a time. The affinity of a gene $a(g)$, is defined to be the sum of similarity values between gene g and all genes in C_{open}. The algorithm alternates between adding high affinity genes to C_{open}, and removing low affinity genes from C_{open} until no more genes can be added to or be removed from it. The C_{open} is closed and forms a new cluster. The genes which are assigned to pervious clusters can be still added to the next cluster. The algorithm iterates until all genes have been assigned to clusters.

2.2 MSSC Algorithm

In our approach, the fuzzy c-means algorithm is to cluster the microarray matrix and the text matrix individually. However, the limitation of many ways is to calculate matrix distances for soft clustering results. Hence, the correlation coefficient in statistics is adopted to calculate the distance of the matrices between the microarray data and literatures. The procedural of MSSC algorithm is in detail as follows:

Algorithm: Multi-Source Soft Clustering (MSSC)
Input: microarray data and biomedical literatures
Output: the final clustering result
Method:
Step 1: Parser the literatures and extract the meaningful terms by tagging, stemming, setting the high frequency threshold and the lower frequency threshold, and filtering out the stop-words.
Step 2: Construct the gene×term matrix, and the element g_{ij} in the matrix is represented as follows:

$$g_{ij} = \left\{ \frac{1}{N_i}\sum_{k=1}^{N_i} \frac{tf_{ij} \times \log \frac{N}{df_j}}{N_{kg}} \right\}$$

Step 3: Normalize the matrices and construct the correlation matrix by the results of fuzzy c-means algorithm. The distance of gene i and gene j was calculated by the probability vectors of the genes with N dimensions.

$$D_{ij} = \frac{\sum V_i V_j - \frac{\sum V_i \sum V_j}{N}}{\sqrt{\left(\sum V_i^2 - \frac{(\sum V_i)^2}{N}\right) \times \left(\sum V_j^2 - \frac{(\sum V_j)^2}{N}\right)}}$$

Step 4: Combine the micorarray matrix M and the literature matrix L by multiplying the individual weight.

$$\text{Matrix}_C = \omega \times \text{Matrix}_M + (1\text{-}\omega) \times \text{Matrix}_L$$

Step 5: Run the soft CAST algorithm for the integrated matrix.

2.3 Matrix Distance

There are many formulas to calculate the matrix distance, such as Euclidean distance, Manhattan distance etc. However, they are hard to deal with the different number of clusters or diverse soft clustering algorithms [15]. Thus, Person correlation coefficient in statistics is adopted to calculate the distance of two diverse results of matrices. The Person correlation coefficient is represented as follows:

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of x_i and y_i, s_x and s_y are the sample standard deviations of x_i and y_i and $-1 \le r_{xy} \le 1$. For the every element x_i of dataset X, it can be denoted a probability vector with n dimensions as follows:

$$V_i = \{P(\frac{c_1}{x_i}), P(\frac{c_2}{x_i}), ..., P(\frac{c_n}{x_i})\}$$

where $\sum_{j=1}^{n} P(\frac{c_j}{x_i}) = 1$, and the probability P_j extracted the results of fuzzy c-means which are the propbility of elements x_i in the cluster c_j. The similarity distance of gene i and gene j is represented as follows:

$$D_{ij} = \frac{\sum V_i V_j - \frac{\sum V_i \sum V_j}{N}}{\sqrt{\left(\sum V_i^2 - \frac{(\sum V_i)^2}{N}\right) \times \left(\sum V_j^2 - \frac{(\sum V_j)^2}{N}\right)}}$$

3 Experimental Results

In this paper, the yeast DNA data is used as the experimental datasets to verify the effect of the MSSC. The reason is that the yeast database is rich and easy to compare the results with the existent biological information. The MSC approach is implemented in Matlab, and the MSSC approach is coding by Matlab and C++.

3.1 Datasets

The microarray data which are generated by Michael Eisen's lab are used as the experimental data to analyze the gene expressions in advance. There are 6221 genes over 80 experimental conditions. As to the literatures, we download 36325 yeast-related abstracts which were published from 1975 to Apr, 2006 from the Saccharomyces Cerevisiae Database (SGD)[5]. In addition, the MIPS Comprehensive Yeast Genome Database (CYGD)[4] provides the main categories of gene functions to verify the results. Further, we choose 978 genes which are related to the transcription function as the function enrichment of the clusters.

3.2 Function Enrichment Tool

In order to evaluate the quality of clustering, the Minkowski Score is adopted [10]. The validity formula is defined as the following equation:

$$\mathrm{MS}(T, C) = ||T - C|| / ||T||$$

where $||T|| = \mathrm{sqrt}(\sum_i \sum_j T_{ij})$.A clustering result for n elements can be represented by an n×n matrix, C. Cij =1 iff xi and xj are in the same cluster, and Cij =0 otherwise. T contains the actual values of genes and represents a reference. Minkowski Score is the normalized distance between the two matrices, C and T. Thus, the smaller the score is, the better the solution will be. Besides, FuncAssociate is used to evaluate the significant level of clusters by the terms from the gene ontology (GO) database [3]. By using the p-value in statistic, the program ranks the significant clusters to enhance the biological meaning. The p-value is one-sided single hypothesis between attribute and query based on Fisher's exact test. The value of $-\log_{10}$ in the experiment is larger and the significant level is higher.

3.3 Results

Table 1 shows the Minkowski score of the clustering results of various databases and methods. From the comparison of the Minkowski score, two results can be induced as follow: First, multi-sources raise the accuracy effectively than the single dataset. Second, MSSC is more accurate than MSC in the yeast dataset. Besides, it deserves to be motioned that the difference of microarray data and literature data results from the hard clustering, k-means. It reveals that heterogonous data may have different fuzzy intensity. In the other paper, our experimental result shows that soft clustering methods are necessary by adjusting the fuzzy exponents of two datasets while the property of clusters is overlapping.

Table 1. Five Clustering Results

Method	Dataset	Minkowski Score
k-means	Microarray Data	1.197
	Literaure Data	1.318
	Multi-Source	1.132
MSC	Multi-Source	1.118
MSSC	Multi-Source	1.006

Table 2 reveals that function enrichment analysis can enhance the significant clusters, especially the unknown dataset. The p-value displays the statistic significant of the number of genes within the clusters in comparison to the whole genome. In the transcription function of cluster, the related genes in MSSC are enriched more and the p-value is higher. Furthermore, it is verified that many genes involve multi-functions in the cluster. It can find out more genes that have common functions and similar reactions in the dataset.

Table 2. Function enrichment of the cluster

Enriched Functional Category (GO attribute)	MSC		MSSC	
	# genes	p-value ($-\log_{10}$)	# genes	p-value ($-\log_{10}$)
ribosome biogenesis	26	18	31	32
transcription	25	7	74	27
RNA metabolism	28	8	45	15
rRNA processing	21	21	31	37
transcription from RNA polymerase II promoter	44	19	56	20
transcription regulator activity	11	7	45	16

4 Conclusion

In the paper, a framework of soft clustering algorithms and multi-sources is presented to aim at integrating heterogonous data. The results of MSSC in the yeast dataset indicate that the method can generalize better quality clusters and provide the significant GO attribute of clusters. In addition, the fuzzy exponent and the weight parameter supply users a pivot to adjust the clustering rules based on the property of the actual datasets.

Clustering multi-sources is a new and promising research issue. There are several aspects for the future research. First, it could add more different databases to enhance the accuracy of the gene expressions. Second, it could supply automatically the best solution without modifying the parameter of input by validity techniques. Finally, the concept could extend to mine the incremental databases and apply in other domain.

Acknowledgement

This research was supported in part by the National Science Council of Republic of China under Grant No. NSC 95-2520-S-024-002.

References:

[1] Ben-Dor, A., & Yakhini, Z., Clustering gene expression patterns. *Journal of Computational Biology*, Vol.6, 1999, pp.281-297.

[2] Berkhin, P., Survey of Clustering Data Mining Techniques. *Technical Report*, Accrue Software, San Jose, CA, 2002.

[3]Berriz, G.F., King, O.D., Bryant, B., Sander, C. & Roth, F. P., Characterizing gene sets with FuncAssociate. *Bioinformatics*, Vol.19, No.18, 2003, pp.2502-2504, from http://llama.med.harvard.edu/cgi/func/funcassociate

[4] Chaussabel, D., & Sher, A., Mining Microarray Expression Data by Literature Profiling. *Genome Biology*, Vol.3, No.10, 2002, Research0055. Retrieved April 26, from http://mips.gsf.de/genre/proj/yeast/

[5] Dolinski, K., & Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engelm, S.R., Fisk, D.G. et al., Saccharomyces Genome Database, 2004, Retrieved Apr., 2006, from http://www.yeastgenome.org/

[6] Eisen, M.B., & Brown, P.O., DNA arrays for analysis of gene expression. *Methods in Enzymology*, Vol.303, 1999, pp.179-205, from http://rana.lbl.gov/index.htm

[7] Glenisson, P., Mathys, J. & Moor, B.D., Meta-clustering of gene expression data and literature-based information. *Journal of ACM Special Interest of Group on Knowledge Discovery and Data Mining Explorations*, Vol.5, No.2, 2004, pp.101-112.

[8] Han, J., & Kamber, M., *Data Mining Concepts and Techniques*. CA：Morgan Kaufmann, 2001.

[9] Hu, X., Integration of cluster ensemble and text summarization for gene expression analysis. *Proceedings of the 4th IEEE International Symposium on Bioinformatics and BioEngineering*, Taichung, Taiwan, 2004, pp.251-258.

[10] Hu, X., & Yoo, Illhoi., Cluster ensemble and its application in gene expression analysis. *Proceedings of the 2nd Asia-Pacific Bioinformatics Conference*, Dunedin, New Zealand, 2004, pp.297-302.

[11]Masys, D. R., Linking microarray data to the literature. *Nat Genet*, Vol.28, No.1, 2001, pp.9-10.

[12]Raychaudhuri, S., Jeffrey, T., Chang, I.F. & Altman, R.B.(2003).The computational analysis of scientific literature to define and recognize gene expression clusters. *Nucleic Acids Research*, Vol.31, No.15, 2003, pp.4553-4560.

[13] Yang, C., Zeng, E., Li, T. & Narasimhan, G., Clustering genes using gene expression and text literature data. *Proceedings of the 4th International IEEE Computer Society Computational Systems Bioinformatics Conference*, Stanford, CA, USA, 2005, pp.329-340.

[14] Yang, C., Zeng, E., Li, T. & Narasimhan, G., A knowledge-driven method to evaluate multi-source clustering. *Proceedings of the Parallel and Distributed Processing and Applications*, Nanjing, China, 2005, pp.192-202.

[15] Zhao, Y., & Karypis, G., Soft clustering criterion functions for parational document clustering. *CSE/UMN technical report*, 2004, TR-04-022.