ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

IJARMATE
Your ulti-MATE Research Paper !!!

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE) Vol. IV, Issue VII, July 2018

EXTRACTION OF BIOLOGICAL KNOWLEDGE BY CLUSTERING DATA MINING TECHNIQUES

P. Lakshmi [1] , Dr.K.Nachimuthu [2]

1. Research Scholar, Department of Computer Science, Jairams Arts and Science College, Karur.

2. Research Advisor, Department of Computer Science, Jairams Arts and Science College, Karur.

Abstract: Bioinformatics and computational science include the utilization of systems including connected arithmetic, informatics, measurements, software engineering, artificial insight, science, and organic chemistry to take care of natural issues for the most part on the atomic level. The center standard of these systems is us-ing registering assets keeping in mind the end goal to tackle issues on sizes of extent unreasonably awesome for human insight. The exploration in computational biol-ogy frequently covers with frameworks science. Real research efforts in this field incorporate grouping arrangement, quality discovering, genome get together, protein struc-ture arrangement, protein structure expectation, forecast of quality articulation and protein-protein associations, and the displaying of development. The colossal measure of information engaged with these examination fields makes the use of information min-ing strategies exceptionally encouraging. These procedures, beginning from numerous sources, for example, the aftereffects of high throughput tests or clinical records, goes for unveiling already obscure information and connections.

1.Introduction

Different data sources became available in recent years. For example, DNA microarray experiments generate thousands of gene expression mea-surements and provide a simple way for collecting huge amounts of data in a short time. They are used to collect information from tissue and cell samples regarding gene expression differences. Compared with traditional tumor diagnostic methods, based mainly on the morphological appearance of the tumor, methods relying on gene expression profiles are more objective, accurate, and reliable .

Clustering is a useful exploratory technique for gene expression data as it groups similar objects together and allows the biologist to identify po-tentially meaningful relationships between the genes. The genes belonging to the same cluster are typically involved in related functions and are fre-quently co-regulated. Thus, grouping similar genes can provide a way to understand functions of genes for which information has not been previously available. Another employment of clustering algorithms is to identify group of redundant genes and then select only a representative for each group to perform a dimensionality reduction.

Although these techniques retrieve good models of biological processes, they are strong dependent by the data employed in the experimental analysis. Moreover, in many biomedical application a background knowledge is not available. Thus, text mining methods applied on published research papers may provide powerful tools to validate the experimental results obtained by other data analysis studies. Initially, analyzing and extracting relevant and useful information from research papers was manually performed by molecular biologists. In last years summarization approaches allow facing with this problem in an automatic way

2. Clustering

The goal of clustering in microarray technology is to group genes or experiments into clusters according to a similarity measure . For instance, genes that share a similar expression pattern under various conditions may imply co-regulations or relations in functional pathways. Thus, clustering could provide a way to understand function of genes for which information has not been available previously process . Furthermore, clustering can be used as a preprocessing step before a feature selection or a classification algorithm, to restrict the analysis to a specific category or to avoid redundancy by considering only a representative gene for each cluster. Many conven-tional clustering algorithms have been applied or adapted to gene expression data and new algorithms, which specifically address gene ex-pression data, have recently been proposed . In the following, first the main challenges of clustering methods are described. Then, an overview of recent works which applied the clustering to microarray data is presented, and finally a comparison of their main characteristics is provided.

3. Clustering challenges

The main challenges regarding the application of clustering to microarray data are (i) the definition of the appropriate distance between objects, (ii) the choice of the clustering algorithm, and (iii) the evaluation of final re-sults. Especially evaluating the results of clustering is a non-trivial task. Each article justifies a specific evaluation criterion, and in literature many criteria exist, such as measures which evaluate the obtained clusters without knowing the real class of objects (i.e., homogeneity and separation), mea-sures which evaluate the agreement between the obtained clusters and the ground truth, and measures which involve the comparison with biological databases (i.e., GO) to measure the biological homogeneity of clusters. In addition, some works also highlight the problem of giving a user friendly rep-resentation of clustering results. Another evaluation criterion could be the clustering computational complexity, even if an evaluation of the complexity and efficiency of a method is very difficult to perform without resorting to extensive benchmark.

4. Clustering algorithms

The similarity between objects is defined by computing the distance between them. Gene expression values are continuous attributes, for which several distance measures (Euclidean, Manhattan, Chebyshev, etc.) may be com-puted, according to the specific problem. However, such distance functions are not always adequate in capturing correlations among objects because the overall gene expression profile may be more interesting than the individual magnitude of each feature . Other widely used schemes for determining the similarity between genes use the Pearson or Spearman correlation coef-ficients, which measure the similarity between the shapes of two expression patterns. However, they are not robust with respect to outliers. The cosine correlation has proven to be more robust to outliers because it computes the cosine of the angle between the expression gene value vectors. Other kinds of similarity measures include pattern based (which consider also sim-ple linear transformation relationships) and tendency based (which consider synchronous rise and fall of expression levels in a subset of conditions).

Once the distance measure has been defined, the clustering algorithms are divided based on the approach used to form the clusters. A detailed description of clustering algorithms applied to microarray has been provided by [1]. Mainly, they can be grouped in two categories, partitioning and hierarchical algorithms.

Partitioning algorithms. This family of clustering algorithms works similarly to k-means [2]. K-means is one of the simplest and fastest clus-tering algorithms. It takes the number of clusters (k) to be calculated as an input and randomly divides points into k clusters. Then it iteratively calculates the centroid for each cluster and moves each point to the closest cluster. This procedure is repeated until no further points are moved to dif-ferent clusters. Despite its simplicity, k-means has some major drawbacks, such as the sensibility to outliers, the fact that the number of clusters has to be known in advance and that the final results may change in successive runs because the initial clusters are chosen randomly.

Several new clustering algorithms have been proposed to overcome the drawbacks of k-means. For example, the genetic weighted k-means algo-rithm is a hybridization of a genetic algorithm and a weighted k-means algorithm. Each individual is encoded by a partitioning table which uniquely determines a clustering, and genetic operators are employed. Authors show that it performs better than the k-means in terms of the cluster quality and the clustering sensitivity to initial partitions.

In [3] the authors described the application of the fuzzy c-means to microarray data, to overcome the problem that a gene can be associated to more than one cluster. The fuzzy c-means links each gene to all clusters via a real-valued vector of indexes. The values of the components of this vector lie between 0 and 1. For a given gene, an index close to 1 indicates a strong association to the cluster. Inversely, indexes close to 0 indicate the absence of a strong association to the corresponding cluster. The vector of indexes defines thus the membership of a gene with respect to the various clusters. However, also in this approach there is the problem of parameter estimation.

Au et al. [3] proposed the attribute cluster algorithm (ACA), which adopts the idea of the k-means to cluster genes by replacing the distance measure with the interdependence redundancy measure between attributes and the concept of mean with the concept of mode (i.e., the attribute with the highest multiple interdependence redundancy in a group).

Hierarchical algorithms. Hierarchical clustering typically uses a pro-gressive combination (or division) of elements that are most similar (or di□er-ent). The result is plotted as a dendrogram that represents the clusters and relations between the clusters. Genes or experiments are grouped together to form clusters and clusters are grouped together by an inter-cluster distance to make a higher level cluster. Hierarchical clustering algorithms can be fur-ther divided into agglomerative approaches and divisive approaches based on how the hierarchical dendrogram is formed. Agglomerative algorithms (bottom-up approach) initially regard each data object as an individual clus-ter, and at each step, merge the closest pair of clusters until all the groups are merged into one. Divisive algorithms (top-down approach) starts with one cluster containing all the data objects, and at each step splits a cluster until only singleton clusters of individual objects remain. For example, Eisen et al. [4] applied an agglomerative algorithm called UPGMA (Unweighted Pair Group Method with Arithmetic Mean) and adopted a method to graph-ically represent the clustered data set, while Alon et al. [5] split the genes through a divisive approach, called the deterministic-annealing algorithm.

The authors applied a density-based hierarchical clustering method (DHC) on two datasets for which the true partition is known. DHC is developed based on the notions of density and attraction of data objects. The basic idea is to consider a cluster as a high-dimensional dense area, where data objects are attracted with each other. At the core part of the dense area, objects are crowded closely with each other, and thus have high density. Objects at the peripheral area of the cluster are relatively sparsely distributed, and are attracted to the core part of the dense area. Once the density and attraction of data objects are defined, DHC organizes the cluster structure of the data set in two-level hierarchical structures, one attraction tree and one density tree. However, to compute the density of data objects, DHC calculates the distance between each pair of data objects in the data set, which makes DHC not effiecient. Furthermore, two global parameters are used in DHC to control the splitting process of dense areas. Therefore, DHC does not escape from the typical difficulty to determine the appropriate value of parameters.

5. Comparison for clustering methods

Richards et al. 121provided a useful comparison of several recent cluster-ing algorithms by concluding that k-means is still one of the best clustering method because it is fast, does not require parallelization, and produces clus-ters with slightly high levels of GO enrichment. Despite this consideration, the hierarchical clustering algorithms are the most used in biological studies. The main advantage of hierarchical clustering is that it not only groups to-gether genes with similar expression pattern but also provides a natural way to graphically represent the data set . The graphic representation allows users to obtain an initial impression of the distribution of data. However, the conventional agglomerative approach su□ers from a lack of robustness because a small perturbation of the data set may greatly change the struc-ture of the hierarchical dendrogram. Another drawback of the hierarchical approach is its high computational complexity.

In general, microarray data are clustered based on the continuous ex-pression values of genes. However, when additional information is available (e.g., biological knowledge or clinical information), it may be beneficial to exploit it to improve cluster quality . Clinical information can be used to build models for the prediction of tumor progression. For example Wang et al. [7] used epigenetic data to determine tumor progression in cancer, and Bushel et al. [8] presented a method to incorporate phenotypic data about the samples.

Au et al. [9] presented a particular validation technique for clustering. They selected a subset of top genes from each obtained cluster to make up a gene pool, and then they run classification experiments on the selected genes to see whether or not the results are backed by the ground truth and which method performs the best. Thus, they exploit class information on samples to validate the results of gene clustering. The good accuracy reached by selecting few genes from the clusters reveals that the good diagnostic information existing in a small set of genes can be effectively selected by the algorithm. It is an interesting new way of clustering validation, by integrating clustering and feature selection.

6. Biclustering

Due to the high complexity of microarray data, in last years the scientists focused their attention on biclustering algorithms. The notion of biclustering was first introduced in [10] to describe simultaneous grouping of both row and column subsets in a data matrix. It tries to overcome some limitations of traditional clustering methods. For example, a limitation of traditional clustering is that gene or an experimental condition can be assigned to only one cluster. Furthermore, all genes and conditions have to be assigned to clusters. However, biologically a gene or a sample could participate in multi-ple biological pathways, and a cellular process is generally active only under a subset of genes or experimental conditions. A biclustering scheme that pro-duces gene and sample clusters simultaneously can model the situation where a gene (or a sample) is involved in several biological functions. Furthermore, a biclustering model can avoid those noise genes that are not active in any experimental condition.

Biclustering of microarray data was first introduced in . They defined a residual score to search for submatrices as biclusters. This is a heuristic method and can not model the cases where two biclusters overlap with each other. Segal et al. [11] proposed a modified version of one-way clustering using a Bayesian model in which genes can belong to multiple clusters or none of the clusters. But it can not simultaneously cluster conditions/samples. Bergmann et al. [12] introduced the iterative signature algorithm (ISA), which searches bicluster modules iteratively based on two pre-determined thresholds. ISA can identify multiple biclusters, but is highly sensitive to the threshold values and tends to select a strong bicluster many times. Gu and Liu (2008) proposed a biclustering algorithm based on Bayesian model. The statistical inference of the data distribution is performed by a Gibbs sampling procedure. This algorithm has been applied to the yeast expression data, observing that majority of founded biclusters are supported by significant biological evidences, such as enrichments of gene functions and transcription factor binding sites in the corresponding promoter sequences.

7. New trends and applications

In the last years many studies were addressed to integrate microarray data with heterogeneous information. Since microarray experiments present few samples, the accuracy of the hypotheses extracted by means of data mining approaches could be low. Using different sources of information (e.g., ontolo-gies, functional data, published literature), the biological conclusions achieve improvements in specificity. For example, multiple gene expression data sets and diverse genomic data can be integrated by computational methods to create an integrated picture of functional relationships between genes. These integrated data can then be used to predict biological functions or to aid in understanding of protein regulations and biological networks modeling .

For feature selection approaches some works integrate Gene Ontology (GO) in the computation of most relevant genes. For example in [13] the authors proposed a method that combines the discriminative power of each gene using a traditional filtering method with the discriminative values of GO terms. Moreover, redundancy is eliminated using the ontology annotations. The results show an improvement of classification performance using fewer genes than the traditional filter methods.The analysis of published literature on some specific topic could improve the results on DNA microarray data. With microarray experiments, hun-dreds of genes can be identified as relevant to the studied phenomenon by means of feature selection approaches. The interpretation of these gene lists is challenging as, for a single gene, there can be hundreds or even thou-sands of articles pertaining to the gene’s function. Text-mining can alleviate this complication by revealing the associations between the genes that are apparent from literature.

Unfortunately, current works are focused on keyword search and abstract evaluation that limit the extraction of biological results done in previous studies, and requires the researchers to further filter the results.The interpretations of microarray results can be improved by using ontolo-gies such as MESH or GO . For example, GOEAST is a web-based user friendly tool, which applies appropriate statistical methods to identify significantly enriched GO terms among a given list of genes extracted by gene expression analysis.Clustering is usually considered as an unsupervised learning approach because no a priori knowledge is assumed at the beginning of the process. However, in the case of gene expression data, some prior knowledge is often available (i.e., some genes are known to be functional related). Thus, inte-grating such knowledge can improve the clustering results. In recent years, some semi-supervised clustering methods have been proposed so that user-provided constraints or sample labels can be included in the analysis. For example, in [14]the authors proposed a semi-supervised clustering method called GO fuzzy c-means, which enables the simultaneous use of biological knowledge and gene expression data. The method is based on the fuzzy c-means clustering algorithm and utilizes the Gene Ontology annotations as prior knowledge to guide the process of grouping functionally related genes. By following the approach of using prior biological knowledge for the fuzzy c-means algorithm, other clustering algorithms such as hierarchical and k-means can be adapted to use prior biological knowledge as well.

Conclusion

Analyzing different data sources in order to extract relevant information is a fundamental task in the bioinformatics domain in order to understand bio-logical and biomedical processes. The aim of the thesis work was to provide data mining techniques in order to extract biological knowledge from differentdata sources (i.e., microarray data and document collection). Different data mining techniques were exploited with different aims. Feature selection analysis is a well-known approach to identify relevant genes for biological investigation (e.g., tumor diseases). Feature selection techniques have proved to be helpful in tumor classification and in understanding the genes related to clinical situations. Similarity measures and clustering approaches are pow-erful analyses to identify set of genes which have a similar behavior under different experimental conditions. Finally, summarization techniques allow to extract and present relevant information stored in documents which can be used for the validation of data analysis results.

References

1.R. Shamir and R. Sharan. Algorithmic approaches to clustering gene expression data. *Current Topics in Computational Biology*, 2002.

2.J. Macqueen. Some methods for classification and analysis of multi-variate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1967, pages 281–297, 1967.

3.W.H. Au, K.C.C. Chan, A.K.C. Wong, and Y. Wang. Attribute clus-tering for grouping, selection, and classification of gene expression data. *IEEE/ACM transactions on computational biology and bioinformatics*, pages 83–101, 2005.

4.M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863, 1998

.5. U.Alon, N. Barkai, DA Notterman, K. Gish, S. Ybarra, D. Mack, and AJ Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745, 1999.

6.A.L. Richards, P. Holmans, M.C. O'Donovan, M.J. Owen, and L. Jones. comparison of four clustering methods for brain expression microar-ray data. *BMC bioinformatics*, 9(1):490, 2008

7.Z. Wang, P. Yan, D. Potter, C. Eng, T.H. Huang, and S. Lin. Her-itable clustering and pathway discovery in breast cancer integrating epigenetic and phenotypic data. *BMC bioinformatics*, 8(1):38, 2007

8.P.R. Bushel, R.D. Wolfinger, and G. Gibson. Simultaneous clustering of gene expression data with clinical chemistry and pathological evalu-ations reveals phenotypic prototypes. *BMC Systems Biology*, 1(1):15, 2007.

9. W.H. Au, K.C.C. Chan, A.K.C. Wong, and Y. Wang. Attribute clus-tering for grouping, selection, and classification of gene expression data. *IEEE/ACM transactions on computational biology and bioinformatics*, pages 83–101, 2005

10.JA Hartigan. Direct clustering of a data matrix. *Journal of the Amer-ican Statistical Association*, pages 123–129, 1972.

11.E. Segal, A. Battle, and D. Koller. Decomposing gene expression into cellular processes. In *Biocomputing 2003: Proceedings of the Pacific Symposium Hawaii, USA 3-7 January 2003*, page 89. World Scientific Pub Co Inc, 2002

12.J. Qi and J. Tang. Integrating gene ontology into discriminative powers of genes for feature selection in microarray data. In *Proceedings of the 2007 ACM symposium on Applied computing*, page 434. ACM, 2007

13.L. Tari, C. Baral, and S. Kim. Fuzzy c-means clustering with prior biological knowledge. *Journal of Biomedical Informatics*, 42(1):74–81, 2009.