International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 3 Issue: 12
ISSN: 2454-4248
440 – 442

# Importance of Similarity Measure in Gene Expression Data-A Survey

Tunga Arundhathi
Assistant Professor, Department of CS&IT, MANUU Hyderabad

***Abstract:*** The usage of data mining techniques in research fields of computational biology include gene finding, genome assembly , prediction of gene expression etc, are very promising because the large amount of data is involved in these research fields. These techniques aims that to disclose the unknown knowledge and relationships. Different data sources are available one such as DNA Micro Array is the technology which enables the researchers to investigate and address issues which are non traceable. DNA Micro Array experiments generates thousands of gene expression measurements and provide a simple way for collecting huge amounts of data in short time. Micro array data analysis allows identifying the most relevant genes for a target disease and group of genes with similar patterns under different experimental conditions.Clustering methods are widely used on gene expression data to categorize genes with similar expression profiles. The goal of clustering in micro array technology is to group genes or experiments into clusters according to a similarity measure. In this paper we introduce the concept of micro Array technology, clustering on gene expression data and survey on similarity measure. Finally we conclude this paper promising that similarity measure plays an important role on gene expression data while using one of the data mining techniques is clustering.

***Keywords****: Similarity Measure, DNA Micro Array, Clustering, gene expression data.*

*****

## 1. Introduction

### 1.1. Introduction to DNA MicroArray Technology

DNA Microarray analysis is one of the fastest growing new technologies in the field of genetic research. Scientists are using DNA microarrays to investigate everything from cancer to pest control. Micro Array technologies as a whole provide new tools that transform the way scientific experiments are carried out. The Advantage of micro array technologies compared with traditional methods in one scale[2].Compared with the traditional approach in genomic research ,DNA micro array is a new technology to investigate the expression levels of thousands of genes simultaneously. The data are collected either experiments in time series during a biological process or experiments of different tissue samples,.[1].DNA micro array (DNA chip or biochip) is a collection of microscopic DNA spots attached to asolid surface. DNA Micro arrays are used to measure the expression levels of large no.of genes simultaneously. Each DNA spot contains picomoles of a specific DNA sequence known as probes or oligos. Micro arrays are microscope slides that contain an ordered series of samples (DNA,RNA,protein,tissues).The type of micro array depends upon the material placed onto the slideDNA ,DNA micro Array;RNA RNA micro Array;Protein Protein Micro Array.

Since the samples are arranged in an ordered manner. Data obtained from the micro array can be traced back to any of the samples.this means that genes on the microarray are addressible.the no.of ordered samples on a micro array can number into the hundreds of thousands. The typical micro aaray contains several thousands of addressable genes.

The most commonly used microarray is the DNA microarray. The DNA spotted onto the slides can be chemically synthesized long oligonucle or enzymatically generated PCR products. The slides contain chemically reactive groups that help to stabilize the DNA onto the slide either by covalent bonds or electrostatic interactions. An alternative technology allows the DNA to be synthesized directly onto the slides itself by a photolithographic process.

DNA micro arrays are used to determine

1. The expression levels of genes commonly termed expression profiling.
2. The sequence of genes in a sample commonly termed minisequencing for short nucleotide reads and mutation or SNP analysis for single nucleotide reads.

### 1.2. Gene Expression Data

Genomics is a way to study many genes thousands of genes or even every gene in an organism all at once. With a few expectations, every cell in our body contains copies of each of our 20,000 genes. Different genes are turned on or off in different cell types.A microarray experiment typically assesses a large number of DNA sequences (genes, cDNA clones, or expressed sequence tags [ESTs]) under multiple conditions.

Gene expression data is usually represented as nXm matrix where n is the number of genes and m is the number of time points or samples. Micro Array features or gene transcripts are the rows of the expression matrix and are represented as vectors. Gene Expression data sets are comprised of gene expression levels over time points. Clusters are generated by clustering algorithms that use a data representation as an input. The way the gene expression data is represented, whether it be a graph,matrix,or vector may ease the computation for the problem on hand.

### 1.3. Introduction to clustering techniques.

 ISSN: 2454-4248
440 – 442

Clustering one of the data mining technique has been used for decades in many fields such as image processing, data mining and artificial intelligence[5], and in recent years has benefited microarray gene expression data analysis in genomic research[6].The goal of clustering in microarray technology is to group genes or experiments into clusters according to the a similarity measure. In General to study or design a clustering analysis for an application, need to consider issues like the measurement of the similarity and dissimilarity and the clustering validations. Clustering can be used as a pre-processing step before a feature section 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 conventional clustering algorithms have been applied or adapted to gene expression data [7,8,9] and new algorithms which specifically address gene expression data. The application of clustering to microarray data is the definition of the appropriate distance between objects and the choice of the clustering algorithm and the evaluation of results.

## 2. Related work.

In [10], 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 computed, according to the specific problem. the similarity between genes use the Pearson or Spearman correlation coefficients, 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 simple 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.

There were eight similarity and dissimilarity measures listed in [5],namely Minkowski distance, Euclidean distance, City-block distance, Sup distance Mahalanobis distance, Pearson correlation, Point symmetry distance,Cosine similarity which have been widel used in various applications[6]. In Euclidean distance and pearson correlations were claimed to be effective similarity measures for the gene expression data. The two additional measures namely Jackknife correlation and Spearman's rank order correlation were discussed to cope with the situations of outliers and non-Gaussian distributions respectively.

In[11], Statistical methods to determine differential expression under different conditions can give insight into the gene functions, and it is purported that genes which are expressed similarly under different conditions or experimental setups are likely to have related biological functions. However, this type of analysis is difficult owing to the nature of microarray data. Expression data are noisy and in many cases unreliable. Many factors may affect the experiment and measurements, thus obscuring signals that might indicate relations between genes. In the absence of precise measures for assessing the significance of similarity based on expression profiles, it is not clear whether genes are indeed truly co-regulated or are functionally linked even when they seem to be similarly expressed. Moreover, the choice of the metric can have a great impact on the analysis, e.g. when clustering genes based on microarray data in search for coordinated groups of co-expressed genes. Indeed, it is well-known that different representations and distance measures can have significant effect on the quality of the clustering results, as most clustering algorithms rely directly on pair wise distances or similarities between instances. This includes k-means, pair wise clustering (also called hierarchical clustering) and spectral clustering algorithms. Therefore, better pair wise measures are likely to produce better results, i.e. clusters that better correlate with cellular processes.

The advent of microarray technology has allowed for the large-scale analysis of gene expression profiles and is accompanied with a myriad of possible applications. Microarray analysis has been used to monitor the expression of genes as a cell undergoes a normal physiological process, such as the cell cycle in an attempt to determine the genes involved in this process. It has been used to study differential gene expression patterns under different environmental conditions others have studied the association between different expression profiles and different cellular conditions. Such associations can help in developing assays that are designed to detect different types of cancers based on the expression patterns of genes. In addition, gene knockout experiments followed by microarray assays have been carried out to determine the role of different genes in cellular processes. Statistical methods to determine differential expression under different conditions can give insight into the gene functions, and it is purported that genes which are expressed similarly under different conditions or experimental setups are likely to have related biological functions.

The Expression data are noisy and in many cases unreliable. Many factors may affect the experiment and measurements, thus obscuring signals that might indicate relations between genes. In the absence of precise measures for assessing the significance of similarity based on expression profiles, it is not clear whether genes are indeed truly co-regulated or are functionally linked even when they seem to be similarly expressed. The choice of the metric can have a great impact on the analysis, e.g. when clustering genes based on microarray data in search for coordinated groups of co-expressed genes. Indeed, it is well-known that

 ISSN: 2454-4248
440 – 442

different representations and distance measures can have significant effect on the quality of the clustering results, as most clustering algorithms rely directly on pair wise distances or similarities between instances. This includes k-means, pair wise clustering (also called hierarchical clustering) and spectral clustering algorithms. Therefore, better pair wise measures are likely to produce better results, i.e. clusters that better correlate with cellular processes.

### 3.1. Methods used for similarity measures.

Our main focus of study is to determine whether two genes have similar expression patterns. Therefore we need to choose an appropriate similarity measure. In [11], the global measures (such as the Euclidean metric, the Pearson correlation and the Spearman rank correlation), statistical measures (Z-score-based), local similarity measures that are based on the dynamic programming algorithm and measures of anti-correlation. Most of these are traditional measures. The new mass-distance (MD) measure is to determine the most effective pair-wise similarity measure.

In[10],The common characteristics of most used clustering approaches applied on microarray data is that they cluster genes only by analyzing their continuous expression values. These approaches are appropriate when there is no information about sample classes and the aim of clustering is to identify a small number of similar expression patterns among samples. However, when additional information is available (e.g., biological knowledge or clinical information), it may be beneficial to exploit it to improve cluster quality.

## 4. IMPORTANCE OF SIMILARITY MEASURE IN GENE EXPRESSION DATA

Besides selecting a clustering algorithm, choosing an appropriate, proximity measure is of greater importance to achieve good clustering results. Distance measures are used for defined relationships between the biological molecules of interest. Clustering algorithms use this relationship in different ways. In general similarity and dissimilarity are important because they are used by a no of data mining techniques, such as clustering, nearest neighbour classification and anomaly detection. The initial data set is not needed once these similarities have been computed. The measures such as correlation and Euclidean distance which are useful for dense data such as time series or two dimensional points, as well as the Jaccard and cosine similarity measures which are useful for sparse data like documents.

## 5. CONCLUSION

There are no guidelines concerning how to choose proximity measures for clusteing micro array data. However pearson is the most used proximity measure. In this concern, the following observations may be considerable.

1. The type of proximity measure should fit the type of data.
2. Many types of dense, continuous data , Euclidean distance are often used.
3. For sparse data , which often consists of asymmetric attributes, we typically employ similarity measure that ignores 0-0 matches.
4. For a pair of complex objects, similarity depends on the number of characteristics they both share.
5. For sparse asymmetric data most objects have only a few of the chatacterstics described by the attributes. The jaccard and extended Jaccard measures are appropriate.

## References

[1] D.M. Dziuda, Data mining for genomics and proteomics: Analysis of gene and Protein expression data, Wiley,2010.

[2] DNA Micro Array Data Analysis , Jarko Tuinala,M.Minna Laine,CSC the Finish IT Centre.

[3] M.D.Schena,R Shalon R.Davis and P.Brown"Quantitative monitoring of gene expression patens with a complementary DNA Microarray"Science,Vol 270,PP 467-470,1995

[4] D Lockhart et al.,"Expression monitoring by Hybridization to high-Density Oligonucleotide Arrays"

[5] R Xu and D.IIWunsch,"survey of clustering algorithems" IEEEtransactons ,Neural Networks, Vol 16,no.3 PP 645-678,2005

[6] D.XJiang C.Tang and A.D Zhang "Expression anysis for gene expression data:A Servey",IEEE trans Knowledge and Data Engineering Vol 16,no.11 PP 1370-1386,2004.

[7] L. Fu and E. Medico. FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. BMC bioinformatics, 8(1):3, 2007.analysis of DNA microarray data. BMC bioinformatics, 8(1):3, 2007.

[8] D. Jiang, J. Pei, and A. Zhang. DHC: A density-based hierarchical clustering method for time series gene expression data. In Proceedings of the 3rd IEEE Symposium on BioInformatics and BioEngineering, page 393. IEEE Computer Society Washington, DC, USA, 2003.

[9] A. Thalamuthu, I. Mukhopadhyay, X. Zheng, and G.C. Tseng. Evaluation and comparison of gene clustering methods in microarray analysis.Bioinformatics, 22(19):2405, 2006.

[10] PhD Thesis Extraction of biological knowledge by means of data mining techniques Author:Alessandro Fiori April 2010 A.A. 2009/2010 III Facolt a di Ingegneria Settore scientifico ING-INF/05

[11] Golan Yona, William Dirks, Shafquat Rahman and David M.Lin 3"Effective Similarity measure for expression profiles" Vol 22 no. 132006, Pages 1616-1622