IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

Semisupervised Based Spatial EM Structure for Microarray Analysis

M.Revathi[1], K.Abinaya[2], R.Geetha[3]

Teaching Fellow, Department of CSE, Anna University (BIT Campus), Trichy, Tamilnadu, India[1]

UG Student, Department of CSE, Anna University (BIT Campus), Trichy, Tamilnadu, India.[2,3]

Abstract: Microarray technology is one of the significant biotechnological means that allows recording the expression levels of thousands of genes concurrently within a quantity of different samples. Among the large amount of genes presented in gene expression data, only little fraction of them is efficient for performing a certain diagnostic test. So implement feature subset selection approach to reduce dimensionality, removing irrelevant data and increase diagnosis accuracy which is able to cluster genes based on their interdependence so as to mine important patterns from the gene expression data using Spatial EM algorithm. It can be used to calculate spatial mean and rank based scatter matrix to extract relevant patterns and further implement classification to diagnosis the diseases. A semi-supervised clustering is shown to be effective for identifying biologically important gene clusters with excellent predictive capability. The experiment results prove that Spatial EM based classification approach provides improved accuracy in diseases diagnosis.

Keywords: Microarray, Gene Expression, Spatial EM, Scatter Matrix, Disease diagnosis.

I. INTRODUCTION

A Microarray is a collection of microscopic DNA symbols. A microarray may contain thousands of spots and each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a gene. A microarray database is a repository containing microarray gene expression data. The key uses of a microarray databases are to store the measurement data, manage a searchable index, and make the data available to other applications for analysis and interpretation. Usually, gene expression data is arranged in a data matrix, where each gene corresponds to one row and each condition to one column. Each element of this matrix represents the expression level of a gene under a specific condition, and is represented by a real number, which is usually the logarithm of the relative abundance of the mRNA of the gene under the specific condition. In this paper, we first analyze existing survey of microarray technology and discuss the basic elements of clustering on gene expression data. In particular, we divide cluster analysis for gene expression data into three categories.

II. RELATED WORK

M. P. Brown, W. N. Grundy...,specifies Support Vector Machine (SVM) which contains mathematical features that build them attractive for gene expressions study and together with their flexibility in choosing a similarity function, sparseness solution when dealing with large datasets, the capacity to handle large feature spaces and the ability to identify the outliers. [1]

S. Bashir, E. M. Carter..,proposes mixture models that include the parameters estimates obtained are the maximum likelihood estimators of the location vectors and the common covariance matrix. In the presence of outliers, these estimators are non-robust. [2]

Y. Chen,Bart H. Jr, X. Dang...,propose a novel statistical depth, the kernelized spatial depth (KSD) to generalizes the spatial depth through positive definite kernels. By choosing a proper kernel, the KSD be able to capture the local structure of a data set while the spatial depth fails. [4]

Y. Chueng...,suggest to learn the model parameters through maximizing a weighted likelihood, which is developed from the likelihood function of inputs with a designable weight. Under a particular weight design, then give out a maximum weighted likelihood (MWL) approach name Rival penalized Expectation-Maximization (RPEM) algorithm, which makes the components in a density mixture compete with other, and the rivals intrinsically penalized with a dynamic control during the learning. [4]

R. Serfling.., present a general approach based on the use of two different matrix-valued scatter measures together to discover interesting features. It provides a general result on construction of affine invariant functional using transformation and illustrate briefly in the context of robust outlyingness functions.[17]

III. EM BASED GENE CLUSTERING

In molecular biology, micro arrays are engaged in monitoring the expression levels of genes simultaneously. Arrays are used in the domain of gene expression, genome mapping, toxicity, pathogen identification and other biological applications.

Clustering is a helpful method for grouping gene expression data. In clustering, similar gene expression data will be grouped simultaneously for identifying relationships between the genes. Clustering of gene

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

expression data is a useful tool for identify co-expressed genes and biologically related grouping of genes, which is an important research area in bioinformatics. Existing methodology includes Expectation Maximization algorithm has been applied for clustering gene expression data. EM is used to initialize the cluster centroids. With these centroids, EM is used to group the data efficiently. Expectation Maximization is used to calculate maximum likelihood estimates given incomplete samples. Silhouette refers to a method of interpretation and validation of cluster. This measure provides a representation of how well each object lies within its cluster.

IV. SPATIAL EM BASED GENE CLUSTERING

A gene-based clustering shall abstract genes as items and samples as features, while sample-based clustering would perceive vice-versa. A third category of clustering type also exists, subspace clustering. Subspace clustering, unlike gene-based or sample-based clustering techniques, is not "global" rather it aims to cluster genes based on their pleasure in any disease, being a part of one or more biological pathways. The proposed methodology includes Spatial EM that can be used to calculate spatial mean and rank based scatter matrix to extract relevant patterns and further implement k-nearest neighbor classification approach to diagnosis the diseases. An important finding is that the proposed semi supervised clustering algorithm is shown to be effective for identifying biologically significant gen clusters with excellent predictive capability.

V. SYSTEM ARCHITECTURE

In a testing phase, first upload the gene datasets and remove the noise in that datasets. Irrelevant Features are considered as noise in this datasets.

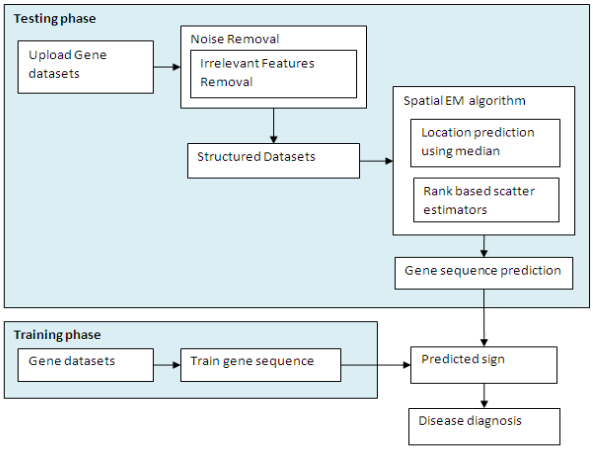

Fig.1.Architecture of semisupervised based spatial EM framework for microarray analysis

In Spatial EM algorithm there are two steps performed to predict the gene sequence. The two steps are Location prediction using median and Rank based scatter estimates. In Location prediction method first find the median value of the datasets and based on that median value we have to cluster the genes. In Rank based scatter estimates method we have to rank the datasets based on how many times that the combination of A,T,C,G present in that datasets. Finally compare the predicted gene sequence with trained gene sequence to diagnosis the disease.

Step by step procedure to diagnosis the disease:

• *Data sets acquisition*

In this module, upload the datasets. The dataset may be microarray dataset and implement the pre-processing steps to eliminate irrelevant data from uploaded datasets. Structured data set can be providing to further module.

• *Median estimation*

In spatial EM, It can analyse coverage of the data before clustering begins and propose an algorithm, which modifies the nearest centroids sorting and the transfer algorithm, of the spatial medians clustering.

• *Rank based scatter*

In this module, can create scatter matrix based on median values that are derived by clustering algorithm. Then construct scatter matrix and reflecting as the within-cluster scatter, the between-cluster scatter and their summation—the total scatter matrix.

• *Disease prediction*

In this module, implement classifier design in semi supervised format. K nearest neighbour classifier allowed to access and provides predicted sign for corresponding diseases such as diabetic, leukemia and so on.

• *Evaluation criteria*

In this module, the performance of the proposed semi supervised algorithm is extensively compared with that of some obtainable supervised and unsupervised gene clustering and gene selection algorithms. The proposed system provide improved accuracy rate in gene classification.

VI. CLUSTERING

Mixture model-based clustering is one of the most popular and successful unsupervised learning approaches. It provides a probabilistic clustering of the data in terms of the fixed posterior probabilities of membership of the mixture components with respect to the clusters. An outright clustering can be subsequently obtained by assigning each observation to the component to which it has the highest fixed posterior probability of belonging. Model-based clustering approaches have a natural way to select the number of clusters based on some criteria, which have the common form of log-likelihood augmented by a model complexity penalty term. In this paper, we deal with robustness of model based clustering. We assume that the number of clusters is known, otherwise, BIC is used. BIC is defined as twice of the log-likelihood minus plogN, where the likelihood is the gaussion based, N is the sample size and p is the number of independent parameters.

For performance assessment, the class labels (ground truth) of training data or testing data are used to construct a confusion matrix (matching matrix). The false positive (rate) (FP/FPR), false negative (rate) (FN/FNR), true positive(TP) and true negative(TN) are computed from the confusion matrix to evaluate the accuracy of clustering methods. We present evaluations of spatial-EM clustering

on the synthetic data and two real data sets. In the simulation experiment, we will see how the number of clusters impacts performance of the EM-based methods. Two real data sets, UCI Wisconsin diagnostic breast cancer data and yeast cell cycle data, are used for comparison of our method and some existing clustering methods.

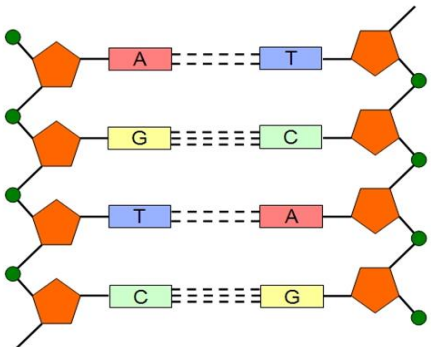

Fig.2. Structure of DNA

- *Spatial EM clustering*

It relies on mainly two factors, namely, determining the relevance of each attribute and increasing the cluster around each related attribute incrementally by adding one attribute after other. Spatial-EM modifies the component estimates on each M-step by Spatial median and rank covariance matrix to gain robustness at the cost of increasing computational burden and losing theoretical tractability.

VII. CLASSIFICATION

KNN is one of the most simple and straight forward learning data mining techniques. It is also called as memory based classification because the training samples require being in the memory at run time. K-Nearest-Neighbour is one of the most frequently used data mining techniques in classification problems. KNN becomes more popular due to its simplicity and relatively high convergence speed. However a main disadvantage of KNN classifiers is the large memory requirement needed to store the whole sample. When the sample is large, reply time on a sequential computer is also large. Despite the memory requirement issue, it is showing better performance in classification problems of various datasets. When dealing with continuous attribute the difference between the attributes is calculated using the Euclidean distance. A major problem when dealing by the Euclidean distance formula is to facilitate the large values frequency swamps the smaller ones. For example, in heart disease records the cholesterol measures ranges between 100 and 190 while the age measure ranges between 40 and 80. So the influence of the cholesterol quantity will be higher than age. To overcome this problem the continuous attributes are normalized so that they have the influence on the distance measure between instances. KNN usually deals with continuous attributes but it can also deal with discrete attributes. When dealing with discrete attributes if the attribute values for the two instances are different so the difference between them is equal to one otherwise it is equal to zero.

VIII. CONCLUSION AND FUTURE WORK

Gene expression data construct by microarray experiment offer tremendous potential for advances in molecular biology and functional genomics. This paper reviewed both classical and recently developed clustering algorithms, which have been realistic to gene expression data, with promising results. The proposed semi-supervised spatial EM clustering algorithm is based on measuring mean values and scatter matrix using the new quantitative measure, whereby redundancy among the attributes is removed. The clusters are then advanced incrementally based on sample categories. The performance of the proposed algorithm is compared with that of existing supervised EM gene selection algorithm with accuracy rate. An important finding is that the proposed semi-supervised clustering algorithm is shown to be effective for identifying biologically important gene clusters with excellent predictive capability.

REFERENCES

1. M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler," knowledge-based analysis of microarray gene expression fidata by using support vector machines,"Proc. Nat. Acad.Sci., vol. 97,no. 1,pp. 262-267, 2000.
2. S. Bashir and E. M. Carter," High breakdown mixture discriminant analysis," J. Multivariate Anal., vol. 93,no. 1, pp. 102-111,2005.
3. N.A. Campbell, "Mixture models and atypical values," Math. Geol., vol. 16, pp .465-477, 1984.
4. Y.Chen, X. Dang, H. Peng, and H. Bart Jr.," Outlier detection with the kernelized spatial depth function," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31,no. 2, pp. 288-305, feb. 2009.
5. Y. chueng,"Maximum weighted likelihood via rival penalized EM for density mixture clustering with automatic model selection," IEEE Tran., Knowl. Data Eng., vol. 17, no. 6,pp. 750-761, Jun. 2005.
6. C. Fraley, and A. Raftery," MCLUST: Software for model-based cluster analysis," J. classification, vol. 16, pp. 297-306, 1999.
7. P. J. Huber,Robust Statistics. New York, NY, USA: Wiley, 1982.
8. C. Liu and D. B. Rubin,"ML estimation of the t distribution using EM and its extension," Stat. Sinia, vol. 5,pp. 19-39, 1995.
9. M. Markaton, "Mixture models, robustness, and the weighted likelihood methodology, "Biometrics, vol. 56, pp.483-486, 2000.
10. R. A. Maronna, R. D. Martin, and V.J.Yohai, Robust Statistics: Theory and Methods. New York, NY, USA: wiley,2006.
11. J. Marden, "Some robust estimates of principal components," stat. probability Lett., vol. 43, pp. 349-359, 1999.
12. G. J. McLachlan and T. Krishan, The EM Algorithm and Extensions. New York, NY,USA: Wiley,1997.
13. G. J. McLachlan and D. Peel, Finite Mixture Models. New York, NY, USA: Wiley, 2000.
14. V. Melnykov and R. Maitra, "Finite mixture models and model-based clustering," Stat. survey., vol.4,pp.80-116,2010.
15. H. Oja. Multivariate Nonparametric Methods with R:An Approach Based on Spatial Signs and Ranks. New York, NY, USA:Springer,2010.
16. C. R. Rao, "Methology based on the L_1-Norm in statistical inference, "Sankhya, Series A, vol. 50,pp. 289-313,1988.
17. R. Serfling," Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardization," J. Nonparametric Stat., vol. 22,pp. 915-936, 2010.
18. D. Tyler, "A distribution-free M-estimator of multivariate scatter," Ann.Stat., vol.15, pp. 234-251, 1987.
19. C. Wallace and D. Dowe, "Minimum message length and Kolmogorov complexity," Comput .J., vol. 42,no. 4,pp. 270-283, 1999.
20. C. F. Wu,"On the convergence properties of the EM algorithm," Annal.Stat., vol.11,pp. 95-103, 1983.
21. J. Oliver, R. Baxter, and C. Wallace," Unsupervised learning using MML, "in proc. 13th Int. Conf. Mach. Learn.,1996,pp. 364-372.

DOI 10.17148/IJARCCE.2016.53121

BIOGRAPHIES

M Revathi is a Teaching Fellow in the Computer Science and Engineering Department, Anna University (BIT Campus), Trichy, Tamilnadu, India. Her area of interest are Data mining, Cloud computing.

K Abinaya is a UG Student in the Computer Science and Engineering Department, Anna University (BIT Campus), Trichy, Tamilnadu, India. Her areas of interests are Data mining, Big Data, Web Designing, Computer networks, Programming in C, C++, JAVA, Php.

R Geetha is a UG Student in the Computer Science and Engineering Department, Anna University (BIT Campus), Trichy, Tamilnadu, India. Her areas of interests are Data mining, Web Designing, Programming in C, .Net, java, HTML.

DOI 10.17148/IJARCCE.2016.53121