‘Gene Micro Array Content Extraction’ An Application Approach Using LFDA, SVM, and Cluster Classification

Thomas Scaria [#1], T Christopher[#2], Gifty Stephen [#3]

[#1] Assistant Professor, Dept Of computer Science, St.Pius X College,Rajapuram,,Kerala,India,
ts.stpius@gmail.com

[#2] Assistant Professor, Dept Of computer Science, Govt.Arts College,Coimbatore,,Tamilnadu,India,
chris_hodcs@gmail.com

[#3] HSST Computer Science, Govt.Higher Secondary,Kottodi,Kerala,India,
giftystephen@gmail.com

Abstract— **Today’s world thrives with data and hence it is necessary to mine the available data, in order to arrive at effective knowledge patterns. Data mining is concerned with extracting knowledge from the available data and so gained knowledge is applied over the processes such as decision making, query processing, and management and so on. The DNA microarray technology enables the biologists to observe the expressions of multiple thousands of genes in parallel fashion. However, processing and gaining knowledge from the voluminous microarray gene data is serious issue. It is necessary for the biologists to retrieve the required data in a reasonable time. The microarray technology enables the researchers to feature reduction of genes using FDA, Gene classification (SVM) and formation of gene clusters within the gene class.**

Keyword - FDA, C Means clustering, SVM and Micro Array

I. INTRODUCTION

Microarray technology is the boon to medical science, which enables the scientific researchers to classify between numerous diseases. Traditionally, the cancer is differentiated with respect to the organ in which the cancer cells grow. The microarray technology enables the researchers to classify between the kinds of cancer by the created patterns of genes. However, the experiments with respect to microarray technology yield a large amount of data and thus the databases are voluminous. Thus, retrieving the user’s data of interest is a serious issue. An effective information retrieval algorithm is required for extracting the required part of the database.

OBJECTIVE OF THE RESEARCH

The main intention of this research work is to present an information retrieval system for microarray gene data and Design and develop an efficient method for gene expression data clustering and classification. Some of the key attributes of an information retrieval system are retrieval speed and accuracy. However, in order to provide considerable speed to the information retrieval system, the dimensionality of the dataset must be deflated without any loss of data. This step is then followed by the classification of gene information with respect to the user’s requirement. After that of classification we try to form gene clusters. To present a novel information retrieval model that is accurate and fast. In most of the recent method, either clustering is concentrated or classification is concentrated. My intension is that, in order to efficiently process the gene data, a clustering and classification algorithm is important

LITERATURE REVIEW

In 2006, Hang X. and Wu X. have presented a system, which diagnoses cancer by means of gene expression. This work utilizes Discrete Cosine Transform (DCT) and Support Vector Machine (SVM). However, the testing samples of this work are limited. In 2009, Sarhan A. has presented a stomach cancer detection system by exploiting Artificial Neural Networks and Discrete Cosine Transform. This work compressed the microarray data by means of Discrete Cosine Transform. The results of this work are satisfactory in terms of specificity, sensitivity and accuracy.

Sheng L. et.al. (2009) presented a system for effective feature selection that relies on LDA techniques, so as to fasten the system. However, this work considers limited training samples. Ruiz R. et.al. (2006) has presented a system gene selection by utilizing the colon and Leukemia datasets by filtering approach. Though this work produces accurate results, it deals with limited number of genes alone.

Xiyi Hang & Fang-XiangWu (2009) have discussed about an approach for cancer diagnosis using gene expression data. Their method symbolized each testing sample as a linear combination of all the training samples. The coefficient vector was acquired by regularized least square. Classification was accomplished by defining discriminating functions from the coefficient vector for each individual category. norm minimization led to sparse solution, and they named the new approach as sparse representation.

Numerical experiments proved that the sparse representation approach matched the best performance accomplished by Support Vector Machines (SVM). Than other classification approach SVM was found to be efficient in performance. Chen Ling and Tu Li (2009) have used Fishers Discriminate Analysis classification method for stream data. It takes less computation and memory space; it is much faster for on-line processing of stream data. It can overcome problem of singular within-class scatter matrix. This method speeds mining process and maintains high accuracy. Number of classes from 10 to 40 were taken and dimensionality was between 10 and 50 were considered. Experiments on synthetic stream data and comparison with other classification methods were performed.

Y.Y Leung and Y.S Hung (2009) proposed combined multiple filters and multiple wrappers. Multiple statistics are used to reduce original set of genes to a manageable size and multiple classifiers and then to iteratively select a small set of genes and used MFMW. This method was used for improving performance and compared with existing method. Three in-one algorithm (weighted voting, k-nearest neighbor and support vector machine) was developed to perform gene selection, sample classification and outlier detection simultaneously. The performance was illustrated with synthetic and real data like leukemia and colon data sets.Thus, very limited works can be cited with respect to information retrieval systems on microarray gene data. However, the shortcomings of the existing works are the absence of dimensionality deflation and most of the works utilized limited size of data.

PradiptaMaji have proposed a supervised attribute clustering algorithm is proposed to find such groups of genes. It directly incorporates the information of sample categories into the attribute clustering process. A new quantitative measure, based on mutual information, is introduced that incorporates the information of sample categories to measure the similarity between attributes. The proposed supervised attribute clustering algorithm is based on measuring the similarity between attributes using the new quantitative measure, whereby redundancy among the attributes is removed. The clusters are then refined incrementally based on sample categories. The performance of the proposed algorithm is compared with that of existing supervised and un supervised gene clustering and gene selection algorithms based on the class separability index and the predictive accuracy of naive bayes classifier, K nearest neighbour rule, and support vector machine on three cancer and two arthritis microarray data sets. The biological significance of the generated clusters is interpreted using the gene ontology. An important finding is that the proposed supervised attribute clustering algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability.

CONCLUSION

The methodology to be followed by this research work is as follows. Initially, the dimensionality of microarray gene data has to be deflated by means of an efficient technique, which is Local Fisher's Discriminant Analysis. Some of the traditional ways of dimensionality reduction techniques are PCA and LDA. However, the major issue of these techniques are the data loss, as it affects the structure of the genomic data.

Thus, we plan to utilize LFDA, which is an improvisation of LDA and FDA. LFDA preserves the structural information of the gene. It is better to deflate the dimensionality of the data, such that the faster response can be expected. After the process of dimensionality deflation, the process of classification has to be done. As this work employs LFDA, the gene information is preserved and thus, the accurate results can be expected. In most of the recent method, either clustering is concentrated or classification is concentrated. My intension is that, in order to efficiently process the gene data, a clustering and classification algorithm is important.The expected outcome of this work is an effective information retrieval system that deals with microarray gene data. The proposed work is expected to be accurate and faster.

REFERENCES

[1] PradiptaMaji and Sushmita Paul, "Rough-Fuzzy Clustering for Grouping Functionally Similar Genes from Microarray Data", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 10, NO. 2, pp. 286- 299,2013

[2] Jianyong Sun, Jonathan M. Garibaldi, and Kim Kenobi, “Robust Bayesian Clustering for Replicated Gene Expression Data", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 9, NO. 5, pp. 1504 - 1512.

[3] Zhiwen Yu, Le Li, Jane You, Hau-San Wong, and Guoqiang Han, " SC3: Triple Spectral Clustering-Based Consensus Clustering Frameworkfor Class Discovery from CancerGene Expression Profiles", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 9, NO. 6, 1751 - 1765, 2012

[4] PradiptaMaji, " Mutual Information-Based Supervised AttributeClustering for Microarray Sample Classification", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1,pp. 127 -140, 2012.

[5] Jaskowiak, P.A, Campello, R.J.G.B. ; Costa, I.G., " Proximity Measures for Clustering Gene Expression Microarray Data: A Validation Methodology and a Comparative Analysis", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol 10 ,No 4, pp: 845 -857, 2013.

[6] H. Causton, J. Quackenbush, and A. Brazma, Microarray GeneExpression Data Analysis: A Beginner's Guide. Wiley-Blackwell,2003.

[7] E. Domany, “Cluster Analysis of Gene Expression Data,”J. Statistical Physics, vol. 110, nos. 3-6, pp. 1117-1139, 2003.

[8] P. Maji and S.K. Pal, Rough-Fuzzy Pattern Recognition: Applicationsin Bioinformatics and Medical Imaging. John Wiley & Sons, Inc.,2012

[9] M.B. Eisen, P.T. Spellman, O. Patrick, and D. Botstein, “ClusterAnalysis and Display of Genome-Wide Expression Patterns,” Proc.Nat’l Academy of Sciences USA, vol. 95, no. 25, pp. 14-863-14-868, 1998.

[10] S. Tavazoie, D. Hughes, M.J. Campbell, R.J. Cho, and G.M.Church, “Systematic Determination of Genetic Network Architecture”,Nature Genetics, vol. 22, no. 3, pp. 281-285, 1999.

[11] A. Brazma and J. Vilo, “Minireview: Gene Expression DataAnalysis,” Federation of European Biochemical Societies Letters,vol. 480, no. 1, pp. 17-24, 2000.

[12] P. D’haeseleer, X. Wen, S. Fuhrman, and R. Somogyi, “Mining theGene Expression Matrix: Inferring Gene Relationships from Large Scale Gene Expression Data,” Proc. Second Int’l Workshop InformationProcessing in Cells and Tissues, pp. 203-212, 1998.

[13] J. Herrero, A. Valencia, and J. Dopazo, “A Hierarchical UnsupervisedGrowing Neural Network for Clustering Gene Expression Patterns,” Bioinformatics, vol. 17, no. 2, pp. 126-136, 2001.

[14] L.J. Heyer, S. Kruglyak, and S. Yooseph, “Exploring ExpressionData: Identification and Analysis of Coexpressed Genes,” Genome Research, vol. 9, no. 11, pp. 1106-1115, 1999.

[15] D. Ghosh and A.M. Chinnaiyan, “Mixture Modelling of GeneExpression Data from Microarray Experiments,” Bioinformatics,vol. 18, no. 2, pp. 275-286, 2002.

[16] G.J. McLachlan, R.W. Bean, and D. Peel, “A Mixture Model-BasedApproach to the Clustering of Microarray Expression Data,” Bioinformatics, vol. 18, no. 3, pp. 413-422, 2002.

[17] K.Y. Yeung, C. Fraley, A. Murua, A.E. Raftery, and W.L. Ruzz,“Model-Based Clustering and Data Transformations for Gene Expression Data,” Bioinformatics, vol. 17, no. 10, pp. 977-987,2001.

[18] M. Dettling and P. Buhlmann, “Supervised Clustering of Genes,”Genome Biology, vol. 3, no. 12, pp. 0069.1-0069.15, 2002.[25] T. Hastie, R. Tibshirani, M.B. Eisen, A. Alizadeh, R. Levy, L.Staudt, W.C. Chan, D. Botstein, and P. Brown, “‘Gene Shaving’ as a Method for Identifying Distinct Sets of Genes with SimilarExpression Patterns,” Genome Biology, vol. 1, no. 2, pp. 1-21, 2000.

[19] T. Hastie, R. Tibshirani, D. Botstein, and P. Brown, “SupervisedHarvesting of Expression Trees,” Genome Biology, vol. 1, pp. 1-12,2001.

[20] Ahmad M Sarhan, 2009, 'Cancer classification based on microarray gene expression data using DCT and ANN', Journal of Theoretical and Applied Information Technology, Vol.6, No.2, pp.208-216.

[21] Hang X. and Wu F.X, 2009, 'Sparse representation for classification of tumors using gene expression data', Journal of Biomedicine and Biotechnology, vol. 2009, pp 1-6.

[22] Lingyan Sheng, Roger Pique-Regi, Shahab Asgharzadeh and Antonio Ortega, 2009, 'Microarray classification using block diagonal linear discriminant analysis with embedded features election', Proceedings of IEEE International conference on Acoustics, Speech and Signal Processing, pp.1757-1760.

[23] Roberto Ruiz, Jose C. Riquelme and Jesus S. Aguilar-Ruiz, 2006, 'Incremental wrapper-based gene selection from microarray data for cancer classification', Pattern Recogntion, vol.39, no.12, pp. 2383-2392.

AUTHOR PROFILE

[#1]Lt.Thomas Scaria is presently working as Assistant professor of Computer Science,at Department Of computer Science,St.Pius X College,Rajapuram-671532,Kerala,India and an active Research Scholar at periyar University Salem.His Area of Interest Includes Data Mining in Bio Sciences and Bio Informatics.He Received his MCA degree from Periyar University in the Year 2003 and M.Phil in Computer Science from Annamalai University in the Year 2008. He Did his M.Sc Applied Psychology from Annamalai University in the Year 2012. He has to Credit 13 Years of teaching and 6 years of research experience.

[#2] Dr.T.Christopher is presently working as Assistant Professor of Computer Science,at Post Graduate and Research Department of Information Technology, Government Arts College(Autonomous),Coimbatore-641018,Tamilnadu,India.He has published 40 research papers in International/National Journals; His area of interest includes knowledge mining and Network security. He has to credit 26 years of teaching and research experience.

[#3] Gifty Stephen is presentaly working as HSST Computer Science at Govt. Higher secondary School Kottody, Previously she worked as Information Manager at Supplyco.She Cleared UGC Net in Computer Science and Just Now started the Research in the area of computer Science.