IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. I (Mar-Apr. 2014), PP 32-36
www.iosrjournals.org

Classification of Micro Array Gene Expression Proposed using Statistical Approaches

Ms. Selva Mary. G [1], Asst. Prof. Sachin M. Bojewar [2]
[1]*PG Scholar, Department of Computer Engineering, Alamuri Ratnamala Institute of Engineering and Technology, Mumbai University, India*
[2]*Asst. Professor, Department of Computer Engineering, Vidyalankar Institute of Engineering and Technology, Mumbai University, India*

Abstract : *Classification analysis of microarray gene expression data has been performed widely to find out the biological features and to differentiate intimately related cell types that usually appear in the diagnosis of cancer. Many algorithms and techniques have been developed for the microarray gene classification process. These developed techniques accomplish microarray gene classification process with the aid of three basic phases namely, dimensionality reduction, feature selection and gene classification. In our previous work, microarray gene classification by statistical analysis approach with Fuzzy Inference System (FIS) was proposed for precise classification of genes to their corresponding gene types. Among various dimensionality reduction techniques, the paper proposed prescribed statistical procedures to efficiently perform the classification process.*

To further substantiate and to analyze the performance, we conduct a comparative study in this work. The comparative study considers two popular dimensionality reduction techniques called Principle Component Analysis (PCA) and Multi-linear Principle Component Analysis (MPCA). The dimensionality reduction techniques replace the proposed statistical approach and perform microarray gene expression data classification. Based on the obtained results, we conduct the performance study over the combination of statistical approach with FIS, LPP with FIS and MPCA with FIS. The study results that the statistical approach with FIS outperforms the classification performance when compared to the other methods.

Keywords: *Classification, k-NN, Micro Array Gene, MPCA, Naive Bayes, PCA, SVM.*

I. INTRODUCTION

Due to wide computation and availability of cost efficient data storages, generous amount of information are being accumulated in the databases. The primary objective of the huge data collection is to determine formerly new patterns and knowledge that aid in efficient decision making process. This necessitates the invention of tools and methods to segregate information that are hidden in such databases and hence the data mining concepts developed. The tradition definition of data mining [25] is "the non-trivial extraction of implicit, formerly unknown and practically beneficial information from data in databases" [1] [2]. It is the fundamental step of Knowledge Discovery in Databases (KDD) [3] in which a defined list of patterns (or models) over the data are generated by deploying the process of computational techniques. Moreover, the inclusion of advancements in the data analysis tools lead to the discovery of more unknown, worth patterns and relationship among the data sets [7-9]. Some of the examples include Statistical models, mathematical algorithms and machine learning methods [4] [5].

There are two broad categories of data mining. They are descriptive data mining and predictive data mining [10]. Some of the examples of descriptive data mining include clustering, association rule mining and sequential pattern mining whereas predictive data mining techniques include classification, regression and deviation detection. In other words, summarizing data and to emphasize their interesting properties is the primary objective of descriptive data mining whereas constructing models to predict the upcoming patterns is the objective of predictive data mining [6]. Among all the predictive data mining techniques, classification plays a significant role in the field of microarray technology. The classification problem is made more in recent days because of concurrent measurement of the expression levels of thousands of genes [14].

Microarray technology is emerged as a robust tool to be used for tracking of genome – wide expression levels of gene [15]. Microarray technologies reveal gene ensembles, the metabolic ways fundamental to the structurally practicable organization of an organ and its physiological function using the analysis of gene expression profiles [16]. They automate the diagnostic process and hence improve accuracy and precision of conventional diagnostic methods. They facilitate thousands of gene expressions [17] [18]. The added advantage of such microarray technology is the ability to classify the cancer types using the micro array gene expression datasets, which ultimately improve the diagnostic measures. Numerous techniques have been proposed so far for the purpose of classifying the cancer types using gene expression datasets [13] [11] [12]. Li-Yeh Chuang et al.

[20] have suggested that support vector machine (SVM) produces equivalent or enhanced results than the neural networks on certain applications. Edmundo Bonilla Huerta et al. [21] integrated Genetic Algorithm (GA) approach with Support Vector Machines (SVM) for the categorization of high dimensional Micro array data. HieuTrung Huynh et al. [22] have exploited feed forward neural network (SLFN) for DNA micro array classification.

PradiptaMajiet al. [23] discussed about the role of information measures like entropy, mutual information, and f-information in dimension reduction of high dimensional micro array data set. Venkateshet al. [24] addressed the problems in conventional dimensionality reduction methods in which gene properties are not accurately presented and hence the important genes from the gene data set are extracted correctly. In the previous work [26], we extracted feature and reduced the dimension of the microarray gene expression based on statistical approach and classification was done using a personalized fuzzy inference system (FIS). This work intends to extend the work by performing a comparative analysis between the proposed statistical approach based dimensionality reduction and the conventional and popular dimensionality reduction techniques such principal component analysis (PCA) and multi-linear principal component analysis (MPCA). The comparative analysis is made in two aspects, one in terms of classification performance and the other in terms of computational complexity. The rest of the paper is organized as follows. Section 2 gives a brief introduction about the statistical approach based dimensionality reduction [26], PCA –based dimensionality reduction and MPCA – based dimensionality reduction.

II. LITERATURE SURVEY

A. Dimension reduction in data mining

Feature selection aims to identify and to remove as much irrelevant and redundant features as possible with respect to the task to be executed. brings some benefits for data mining, such as: an improved predictive accuracy, more compact and easily understood learned knowledge and reduced execution time for algorithms[1]. As a possibility to reduce the number of feature considered by data mining algorithms, in order to make them more efficient, this paper presents a method which uses a combination filter-wrapper[25]. We have used a correlation based filter on the whole set of features, then on relevant subset of features we have applied a wrapper which uses a decision tree classifier for prediction[18]. As a case study we have applied this method on data collected by TERAPERS a system which aims to assist speech therapists on personalized therapy of dyslalia. We have compared the performances obtained both for feature selection by the described method, and for feature selection using only the same wrapper as in first case[25]. We have achieved clearly superior performances for execution time, when we have used for feature selection the combined approach and backward selection as search strategy for wrapper. The positive results obtained for the considered data encourage us to continue our work[2]. We will try to improve these execution times by parallelization of feature selection operations.

Merits:

- Gives better result compared to existing

Demerits:

- The proposed approach is ideal for visualization of high dimensional data.

B. Verdict accuracy of quick reduct algorithm using clustering, classification techniques for Gene expression data

Feature selection (FS) is a process which attempts to select more informative features. This paper studies a feature selection method based on rough set theory[17]. Further K-Means, Fuzzy C-Means (FCM) algorithm have implemented for the reduced feature set without considering class labels. Then the obtained results are compared with the original class labels. Back Propagation Network (BPN) has also been used for classification[7][22]. Then the performance of K-Means, FCM, and BPN are analyzed through the confusion matrix. It is found that the BPN is performing well comparatively [24]. In this paper, Quick reduct algorithm based on rough set theory has been studied for gene expression datasets. The reduced feature set has been used to cluster the data using K-Means and FCM algorithms with considering decision attributes. The performance was evaluate during confusion matrix with positive and negative class values. Further, the selected features with class labels were classified using Back Propagation Network[22]. It was observed that the performance of the BPN is significant.

Merits:

- High Accuracy

Demerits:

- Time complexity

C. A study on effective mining of Association rules from huge databases

This paper [3] provides an overview of techniques that are used to improvise the efficiency of Association Rule Mining (ARM) from huge databases. Owing to the current explosion of information and the accessibility of cheap storage, collecting enormous data has been achievable during the last decades. The ultimate intent of this massive data collection is the utilization of this information to achieve competitive benefits, by determining formerly unidentified patterns in data that can direct the process of decision making. Lately it has been demonstrated that analysis of data with the aid of Online Analytical Processing (OLAP)[6] tools alone is highly tedious; Mining association rules is a prototypical problem as the data are being generated and stored every day in corporate computer database systems. To manage this knowledge, rules have to be pruned and grouped, so that only reasonable numbers of rules have to be inspected and analyzed. Thus an appropriate technique has to be employed to mine the association rule efficiently.

Merits:

- High efficiency

Demerits:

- Time complexity

D. Distributed frequent item sets mining in heterogeneous platforms

Huge amounts of datasets with different sizes are naturally distributed over the network. In this paper we propose a distributed algorithm for frequent item sets generation on heterogeneous clusters and grid environments [5]. In addition to the disparity in the performance and the workload capacity in these environments, other constraints are related to the datasets distribution and their nature, and the middleware structure and overheads. The proposed approach uses a dynamic workload management through a block-based partitioning, and takes into account inherent characteristics of the Apriori algorithm related to the candidate sets generation. The proposed technique greatly enhances the performance and achieves high scalability compared to the existing distributed Apriori-based approaches. This approach is evaluated on large scale datasets distributed over a heterogeneous cluster. Experiments have been conducted on a heterogeneous test bed and show that the proposed algorithm achieves very good performance and high scalability compared to a classical Apriori-based implementation.

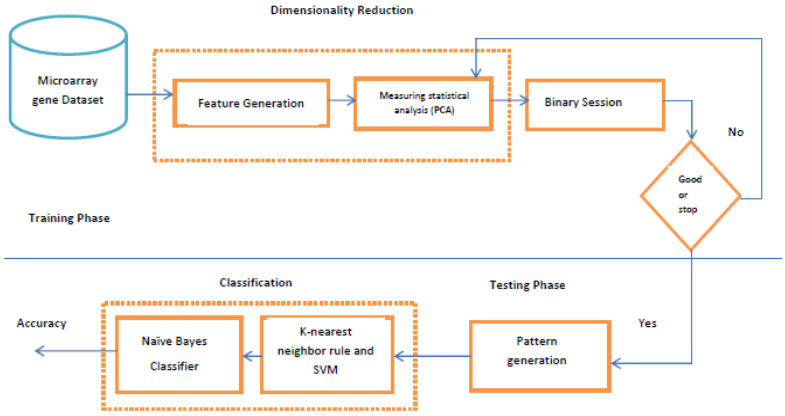
III. EXISTING SYSTEM

In this paper existing system is Fuzzy inference system (FIS) [19]. FIS based method was accuracy efficient, however the reliability is poorer than the proposed method and ultimately, the classification performance is poorer than the proposed method. The classification performance of the fuzzy inference system (FIS) is similar to that of other classifiers, but simpler and easier to interpret. Consequently, the goal is to generate fuzzy rules based on dimensionality reduced data. Hence, fuzzy inference is selected in our approach for classification and the fuzzy rules are utilized to train the fuzzy inference system [28].

IV. PROPOSED SYSTEM

In our proposed method we will classify the Micro Array Gene Expression Data. Features are extracted in the data and extracted features are reduction using PCA, MPCA method. Selected features are passed through binary session. After the binary session, Patterns are generated using statistical approach. Then the features and patterns are passed through the classifiers. Finally Measure Performance metrics like accuracy, sensitivity, specificity, FPR, PPV, NPV, FDR and MCC and concentrated on the computational time.

V. METHODOLOGY



Modules

- Two main Modules
  - Training Phase
    - Load Microarray Gene Dataset
    - Feature Extraction using Dimensionality Reduction (MPCA and PCA)
    - Binary Session
  - Testing Phase
    - Pattern Generation
    - K-nearest neighbor rule and SVM
    - Naïve Bayes Classifier

Finally Measure Performance metrics like accuracy, sensitivity, specificity, FPR, PPV, NPV, FDR and MCC and concentrated on the computational time.

VI. CONCLUSION

This paper studied the performance of the proposed statistical approach based dimensionality reduction in microarray gene classification method over the popular dimensionality reduction and feature extraction methods such as MPCA – based dimensionality reduction and PCA – based dimensionality reduction. Hence, it can be asserted that the proposed dimensionality reduction method is suitable for microarray gene classification as it extracts relevant and less volume of information from the raw expression. Finally Measure Performance metrics like accuracy, sensitivity, specificity, FPR, PPV, NPV, FDR and MCC and concentrated on the computational time.

REFERENCES

[1] Osmar, "Introduction to Data Mining", In: Principles of Knowledge Discovery in Databases, CMPUT690, University of Alberta, Canada, 1999

[2] Kantardzic and Mehmed, "Data Mining: Concepts, Models, Methods, and Algorithms", John Wiley & Sons, 2003

[3] Umarani and Punithavalli, "A Study on Effective Mining of Association Rules from Huge Databases", International Journal of Computer Science and Research, Vol. 1, No. 1, pp. 30-34, 2010

[4] Chieh-Yuan Tsai and Min-Hong Tsai, " A dynamic Web service based data mining process system", In Proceedings of the 5th IEEE International Conference on Computer and Information Technology, pp. 1033-1039, 21- 23 September, 2005

[5] Lamine M. Aouad, Nhien-An Le-Khac and Tahar M. Kechadi, "Distributed Frequent Itemsets Mining in Heterogeneous Platforms", Journal of Engineering, Computing and Architecture, Vol. 1, No. 2, 2007

[6] J. Han and M. Kamber, "Data Mining: Concepts and Techniques. Morgan Kaufman, San Francisco, 2000

[7] Bigus, "Data Mining with Neural Networks", McGraw-Hill, 1996

[8] Klaus Julisch," Data Mining for Intrusion Detection -A Critical Review", In Proceedings of the IBM Research on application of Data Mining in Computer security, Chapter 1 , 2002

[9] Hewen Tang, Wei Fang and Yongsheng Cao, "A simple method of classification with VCL components", In Proceedings of the 21st international CODATA Conference, 2008

[10] Miller, Jason, "Core Privacy: A Problem for Predictive Data Mining." Lessons from the Identity Trail. New York: Oxford University Press, 2009

[11] Yendrapalli, Basnet, Mukkamala and Sung, "Gene Selection for Tumor Classification Using Microarray Gene Expression Data", In Proceedings of the World Congress on Engineering, London, U.K., Vol. 1, 2007

[12] Sandrine Dudoit, Jane Fridlyand and Terence P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data", Journal of the American Statistical Association, Vol. 97, pp. 77-87, 2002

[13] Peterson and Ringner, "Analyzing Tumor Gene Expression Profiles", Artificial Intelligence in Medicine, Vol. 28,No.1, pp. 59-74, 2003

[14] AnandhavalliGauthaman, "Analysis of DNA Microarray Data using Association Rules: A Selective Study", World Academy of Science, Engineering and Technology, Vol.42, pp.12-16, 2008

[15] Chintanu K. Sarmah, SandhyaSamarasinghe, Don Kulasiri and Daniel C., "A Simple Affymetrix Ratio-transformation Method Yields Comparable Expression Level Quantifications with cDNA Data", World Academy of Science, Engineering and Tech,Vol.61,pp.78-83, 2010

[16] Khlopova, Glazko and Glazko, "Differentiation of Gene Expression Profiles Data for Liver and Kidney of Pigs", World Academy of Science, Engineering and Technology, Vol. 55, pp. 267-270, 2009

[17] Ahmad m. Sarhan, "Cancer classification based on microarraygene expression data using dct and ann", Journal of Theoretical and Applied Information Technology, Vol. 6, No. 2, pp. 207-216, 2009

[18] Ying Xu, Victor Olman and Dong Xu, "Minimum Spanning Trees for Gene Expression Data Clustering", Genome Informatics, Vol. 12, pp. 24–33, 2001

[19] LucilaOhno-Machado, StaalVinterbo and Griffin Weber, "Classification of Gene Expression Data Using Fuzzy Logic", Journal of Intelligent & Fuzzy Systems, Vol. 12, No. 1, pp. 19-24, January 2002

[20] Li-Yeh Chuang, Cheng-Hong Yang and Li-Cheng Jin, "Classification Of Multiple Cancer Types Using Fuzzy Support Vector Machines And Outlier Detection Methods", Biomedical Engineering applications, Basis and Communications,Vol.17,pp.300-308, Dec 2005

[21] Edmundo Bonilla Huerta, Beatrice Duval and Jin-Kao Hao, "A hybrid GA/SVM approach for gene selection and classification of micro array data", In Lecture Notes in Computer Science, pp. 34-44, Springer, 2006

[22] HieuTrung Huynh, Jung-JaKimandYonggwan Won, "Classification Study on DNA Micro array with Feed forward Neural Network Trained by Singular Value Decomposition", International Journal of Bio- Science and Bio- Technology Vol.1,No.1,pp.17-24,Dec, 2009

[23] PradiptaMaji and Sankar K. Pal, "Fuzzy–Rough Sets for Information Measures and Selection of Relevant Genes from Micro array Data", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 40, No. 3, pp. 741-752, June 2010
[24] Venkatesh and Thangaraj, "Investigation of Micro Array Gene Expression Using Linear Vector Quantization for Cancer", International Journal on Computer Science and Engineering, Vol. 02, No. 06, pp. 2114-2116, 2010
[25] Brahmadesam Krishna and BaskaranKaliaperumal, "Efficient Genetic-Wrapper Algorithm Based Data Mining for Feature Subset Selection in a Power Quality Pattern Recognition Application", The International Arab Journal of Information Technology, Vol. 8, No. 4, pp. 397-405, October 2011
[26] Tamilselvi, M.; G.M.Kadhar Nawaz, "Classification of Micro Array Gene Expression Data using Statistical Analysis Approach with Personalized Fuzzy Inference System", International Journal of Computer Applications Volume 31– No.1, p.p. 5-12, October 2011
[27] ALL/AML datasets from http://www.broadinstitute.org/cancer/software/genepattern/datasets/.