Classification of Micro Array Gene Expression using k-NN, SVM and Naive Classifiers

Selva Mary. G, Likhesh N. Kolhe, Kanchan D. Patil

Asst. Professor, Alamuri Ratnamala 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.

This paper proposes an effective classification technique that uses Naïve Bayes classifier, k-NN and SVM. The dimensionality reduction of the gene expression dataset is performed by using statistical approaches. From the dimensionality reduced data, the important genes are identified and also features are extracted. The improved classifier is developed for classification and so it is trained using the classifiers. The well-trained classifier is used to the classification of micro array gene expression dataset.

Index Terms- 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].

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].

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. Previous Work

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)[28]. 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 [29].

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].

In the existing system, it was proposed an effective classification technique that uses an enhanced fuzzy inference classifier. The dimensionality reduction of the gene expression dataset is performed by using statistical approaches. From the dimensionality reduced data, the important genes are identified and also the fuzzy rules are generated. The improved classifier is developed for classification and so it is trained using fuzzy rules. The well-trained classifier is used to the classification of micro array gene expression dataset. Fig.1 describes the architecture of the existing system.

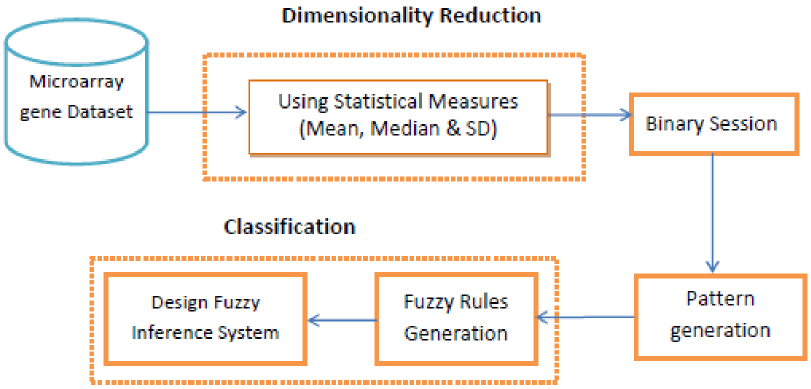

Fig. 1 Existing System Architecture [28]

LIMITATIONS

In this paper, 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 (FIS).

The generation of the Fuzzy rules is very tiresome process and the One of the disadvantages of fuzzy logic is that the rules for it are not very direct. Many experts have proposed rules over the years for this, but there are many of them. It would be impossible to follow all of these rules, since they tend to vary from researcher to researcher. All of the factors of fuzzy logic are given the same importance as long as they are going to be combined.

This causes certain sets of data to become more important than others, where this importance may not necessarily be true. There are many complexities in developing the fuzzy rules. It is also hard to develop a model from a fuzzy system. It requires more fine tuning and simulation before operational. If the rules generated were wrong once then the whole process has to be changed. The use of fuzzy logic for interpretation of gene expression data has not been explored substantially.

Initial investigations suggested that poor quality clusters are formed as a result of the fuzzification of measurements. Estimating the generalizability of these experiments is not easy because computing the quality of clusters continues to be an extremely subjective task, and diverse fuzzification systems have not been tried. The dimensionality reduction of the data has to be also included since the data gene data contains a large amount of values in them.

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.

FIS based method of classification 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 simpler and easier to interpret [28]. Consequently, the goal is to generate fuzzy rules based on dimensionality reduced data. The generation of the Fuzzy rules is very tiresome process and the one of the disadvantages of fuzzy logic is that the rules for it are not very direct. Many experts have proposed rules over the years for this, but there are many of them. It would be impossible to follow all of these rules, since they tend to vary from researcher to researcher. This causes certain sets of data to become more important than others, where this importance may not necessarily be true. There are many complexities in developing the fuzzy rules. It is also hard to develop a model from a fuzzy system. It requires more fine tuning and simulation before operational.

While classifying the gene data type it is more important to have accuracy and reliability rather than time complexities and space complexities. Hence we propose the new method by combining kNN and SVM classifiers to improve the accuracy of classification.

V. METHODOLOGY

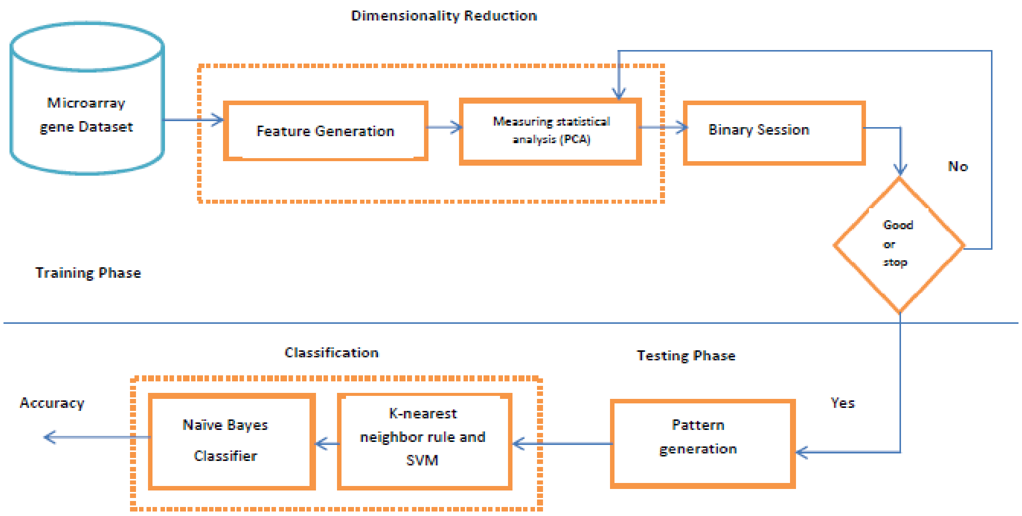

Fig. 2 Proposed System Architecture [29]

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

Training Phase

The Fig. 4 shows the working of the system. In training process, we select all genes from the original data set but except one column and this column of genes is used for testing.

(i) Statistical operations are performed using dimensionality reduction pseudo code

(ii) Then binary session and pattern generation processes are performed based on the result of statistical approach.

(iii) Finally, the fuzzy rules generation process is performed over the generated patterns.

In this proposed work, the first process is loading the microarray Gene dataset. Next phase is feature extraction and it is done by using MPCA and PCA. Binary session is carried out by threshold value. The threshold value in the binary session is used to change the values in the matrixes. It is also used to reduce the execution complexity; binary session makes the classification easier in the further gene expressions.

Testing Phase

- **Pattern Generation**

The binary session process gives the results as patterns. In the binary session each binary elements are obtained as the set of matrix with each matrix. In the set of matrix the first element of the each matrix is considered as one gene pattern, the second element is considered as another gene and this process is repeated for the last value of the given matrix.

In our process we used two types of datasets for the gene expressions named AML and ALL. The generated pattern P contains column g from 1:38, whereas ALL dataset contains 1:26 and AML dataset contains 27:38.

- **K-nearest neighbor rule and SVM**

The k-NN rule is used to introduce notation. In the k-NN rule the patterns which we are going to classify are symbolized as vectors in a d-dimensional Euclidean space R^d.

- **Naïve Bayes Classifier**

Naïve Bayesian classifier is a simple probabilistic classifier model which is based on the Bayes Theorem with independence assumptions. The probabilistic model of the classifier is the “independent feature model”. Since the probability model is a unique in nature, Naïve Bayes classifiers could be trained effectively. In several applications parameter estimation for Naïve Bayesian model uses the maximum likelihood method. Naïve Bayes classifier needs only a small amount of training data to calculate the parameter which are needed for the classifications. This is the major advantage over the Naïve Bayes classifier.

ADVANTAGES

We have included the optimization scheme based on PCA so that the data is reduced. So automatically the time and the cost complexities are much reduced. We are using other classifier such as kNN, SVM and Naïve Bayes classifier and also we are combining the classifiers in order to produce best accuracy in the classification process compared to the existing methods. Instead of the complex fuzzy rules and others we are producing simple and efficient classifiers that are supervised and produce results without many complexities.

The dimensionality reduction scheme that is included helps to produce the classification results in an exact way and also minimizes the complexities. The Bayes classifier classifies the data based on the Bayes theorem calculation which refers to the probability of the test data to occur in a particular category.

The SVM classifier classifies the data based on the hyper plane arrangement of the train data and tests the test data by finding the location at which the test data occurs. The kNN classifier classifies the data by comparing the test and the train data in the neighborhood and finding the results. Thus our proposed scheme produces better accuracy compared to the existing algorithm. Also we have compared the performance of the three classifiers and the classification results by combining the classifiers [12].

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 concentrated on the computational time. It can be seen that the proposed technique has good classification accuracy compared to Fuzzy Genetic, Fuzzy Neural Network ProbPCA and PCA classifiers and fuzzy inference classifier (FIS). The results could show that the proposed system performs satisfactorily in classifying the micro array gene expression dataset

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/.

[28] Selva Mary. G, Sachin Bojewar “Classification of Microarray Gene Expression: A Comparative Analysis using Dimensionality Reduction Techniques“ “Pragyanam ‘14” the proceedings of International Conference on Recent Trends in Computer and Electronics Engineering(ICRTCEE), January, 2014.

[29] Selva Mary. G, Sachin Bojewar “Classification of Micro Array Gene Expression Proposed using Statistical Approaches” International Organization of Scientific Research - Journal of Computer Engineering (IOSR-JCE) (e-ISSN: 2278-0661, p- ISSN: 2278-8727, PP 32-36, Volume 16, Issue 2, Ver. I (Mar-Apr. 2014).

AUTHORS

First Author – Selva Mary. G, Asst. Professor, Alamuri Ratnamala Institute of Engineering and Technology, Mumbai University, India

Second Author – Likhesh N. Kolhe, Alamuri Ratnamala Institute of Engineering and Technology, Mumbai University, India

Third Author – Kanchan D. Patil, Alamuri Ratnamala Institute of Engineering and Technology, Mumbai University, India