Liver Cancer Classification Using Principal Component Analysis and Fuzzy Neural Network

Ms. G.Ilakkiya
Research Scholar
School of Computer Studies
RVS College of Arts and Science
Sulur, Coimbatore - 402, TN, India

Mrs.B.Jayanthi
Assistant Professor
School of Computer Studies
RVS College of Arts and Science
Sulur, Coimbatore - 402, TN, India

Abstract

Cancer research is an exciting area in the medical research field. Categorization is momentously essential for cancer identification and treatment. The exact prophecy of unlike tumor categories has immense value in granting better care and harmfulness reduction on the patients. Categorization of patient taster acquire as gene expression profiles has become a problem of predominant analysis in biomedical research in recent years.

In previous techniques, cancer classification relies on the morphological and clinical. The Proposed arrival of the micro array technology has allowed the repeated study of thousands of genes, which encouraged the development in cancer classification using gene expression data. This analysis core on the widely used mixed data mining and machine learning methods for applicable gene selection, which experiments to accurate cancer classification.

After that, we classify the microarray data sets with a fuzzy neural network (FNN) that we proposed in this study. This FNN combines with Principal Component Analysis with vital features of basic fuzzy model self-generation, constraint enhancement, and rule-base popularization and here used the FNN & PCA to well-known gene expression data sets, i.e., the liver cancer data set.

Keywords: ranking query, web database, deep web

1. Introduction

Data mining or knowledge discovery (KDD) is the process of discovering expressive, new correlation identification and developments by shifting through massive amount of data collection in storehouse, using pattern identification methods as well as statistical and mathematical systems. Data mining is considered as the significant citation of implicit, formerly unknown, and possibly useful information from data.

Microarrays [1] [2] are efficient of profiling the gene expression model of thousands of genes in a first attempt of an experiment. Gene expression [3] data can be a precious source for aware the genes and the biological relationships between the genes. It has high dimension, small tasters and the gene selection Feature selection method is much essential to verify the classification accurateness. The dataset used for this work is called Liver cancer dataset which includes thousands gene expression values with its subtypes.

Liver cancer is complex to be analyzed at a previous step due to reduced amount of obvious signs and the methods inner parts of the body. Therefore, new methods for previous detection of liver cancer are now required to verify the status of the liver and treat patients before the weakening of the liver. The liver cancer microarray data sets took from various hospitals and stored it in the database. Data mining known as extracting or “mining” knowledge from massive or enormous amounts of data.

In recent years numerous techniques were proposed in the literature for gene selection and cancer classification. Data mining algorithms are the most extensively used to classify gene expression data, in these divination of the disease plays an important role for cancer classification. DNA micro arrays are also commonly known, as gene chips, DNA chip, or biochip. In which it is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA micro arrays to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. Important knowledge can be extracted from these data by the use of data analysis techniques.

Bioinformatics is the area of biological research, computer science research, artificial intelligence, statistics and mathematics research that recognize and find interesting knowledge related with large-scale databases. Classification is an essential part of biology and thus classification methods acts a vital role in bioinformatics, repeatedly using similarities of structure to infer alike of function.

In this paper, in medical diagnosis such dataset are highly required as medical researchers can improve more information needed for each diagnosis, we need knowledge innovation from clinically collect dataset for liver cancer which is a prolonged disease a major public health challenge in the world. According to international statistics over 170 million liver patients exist in the world and this may rise with in another 2 to 3 years. So we propose an algorithm Principal Component Analysis and fuzzy neural network used for gene selection and accurate liver cancer classification.

2. Related Works

Artificial neural networks have attained considerable recognition rates in cancer classification. Bevilacqua et al., [6] presented a new technique to artificial neural network (ANN) topology optimization that makes use of the multi-objective genetic algorithm with the intention of discovering the most excellent network configuration for the Wisconsin breast cancer database (WBCD) classification problem.

Kermani et al., [7] revealed that by using a hybrid genetic algorithm and neural network (GANN), the feature extraction can be done more efficiently. An additional benefit of augmenting neural network training with a genetic algorithm is that the extracted features using genetic algorithm are clear and perceivable. Even though the authors estimated this approach by using the breast cancer data, the method is proposed to handle any other kind of classification task.

Lipo Wang et al., [8] aimed at finding the smallest set of genes that can ensure highly accurate classification of cancers from microarray data by using supervised machine learning algorithms.

A statistical method for ranking differentially expressed genes was recommended by Broberg [9]. Current methods have evaluated gene selection approaches by utilizing ROC curves calculated by simulation. But, no effort has been made to evaluate selection accuracy as a function of population parameters.

Feng Chu et al., [10] examined the developed Radial Basis Function network in three standard data sets, i.e., the lymphoma data set, the small round blue cell tumors (SRBCT) data set, and the ovarian cancer data set. The results of these three standard data sets find that this network is capable of realizing 100% accuracy with smaller amount of genes than the existing approaches.

Hero [11] put forth a gene selection and ranking with microarray data. For instance, by utilizing gene microarrays, it is now very easy to investigate a person's gene expression profile more than 30,000 genes of the person genome. Signals obtained from gene microarray experimentations can be associated to genetic features underlying disease, improvement, and aging in a population. This has significantly speeded up the gene detection. However, the enormous scale and investigational variability of genomic data makes removal of biologically important genetic information is very challenging. One of the biggest disputes is to recognize the affected genes that are participated in that specific disease based on a gene microarray research. The authors illustrated multi criterion approaches that are proposed for this gene selection and ranking difficulty.

3. Background Study

Hepatocellular carcinoma (HCC or liver cancer) is one of the most dominant and deadly disease in human beings and it is linked with risk factors such as smoking, obesity, cirrhosis, and hepatitis B and C. In specific, over 78 percent of liver cancer patients are associated with cirrhosis.

The liver cancer data set [12] has two classes, i.e., the nontumor liver and HCC. The data set contains 82 HCCs and 74 nontumor livers. The data is randomly divided. In that half of them are training samples and remaining are testing samples. The data set also has some missing values. To fill those missing values K-nearest neighbor method is used.

3.1 Challenges in Cancer Classification

The same feature of the existing gene expression data set is the major challenge. Massive number of unrelated attributes or genes is another issue. Next challenge is from the application domain of cancer taxonomy. Though Accuracy acts as an important factor in cancer categorization, the biological significance is another key factor, as any biological information visible during the process can support in added gene function innovation and other biological studies.

In the previous system, cancer classification methods are initiated to have many disadvantages in their diagnostic efficiency. To overcome those disadvantages in cancer classification, effective methods in compliance with the overall gene expression analysis have been developed. The expression level of genes controls the result to overcome the disadvantages relevant to the hindrance and cancer treatment. The microarray gene data must be used for categorizing with best accuracy using the classifier which method is used to support that task.

The genes used in the expression outline are not useful and many of them are irrelevant. Decreasing the number of genes by feature selection and still holding best class prediction exactness for the classifier is essential in factor of cancer classification. The stress in cancer classification is based on of gene selection and classifier methods. In Earlier methods Furey and colleagues have used GA-FEC as classifier resulting in 90.3% accuracy in prediction. Li and team have made use of combining GA and PSO method to detect genes that can jointly distinguish between the affected cancer and normal classes. It is a randomly determined, supervised pattern recognition process which achieved 91.1% accuracy with it.

3.2 Proposed Techniques

In this paper, in medical diagnosis such dataset are highly required as medical researchers can improve more information needed for each diagnosis, we need knowledge innovation from clinically collect dataset for liver cancer which is a prolonged disease a major public health challenge in the world. According to international statistics over 170 million liver patients exist in the world and this may rise with in another 2 to 3 years. So we propose an algorithm Principal Component Analysis and fuzzy neural network used for gene selection and accurate liver cancer classification.

4. Methodology

Cancer classification depends on microarray gene expressions are a major issue used a t- test-based feature collection approach to choose some important genes from group of genes.

This paper uses enrichment score for ranking the gene and then the classifier is experimented with that data. At last, the classification of gene for cancer detection is evolved. The experiment is performed with the support of liver cancer data set in which the experimental result shows that the proposed method results in better accuracy and use minimum time for classification when compared to the usual methods.

4.1 Fuzzy Neural Network

The researchers classify the microarray data sets by using the FNN algorithm and this FNN put together essential features of fuzzy generalization. By considering the lesser amount of genes needed by the FNN and its elevated accuracy, it is to be concluded that the FNN classifier not only supports biological researchers to differentiate cancers that are complex to be classified using normal clinical techniques, but also helps biological researchers to focus on a minimum number of important genes to explore the relationships among those significant genes and the development of cancers. It is very vital for cancer diagnosis and treatment to perfectly recognize the site of foundation of a tumor. With the materialization and huge progression of DNA microarray approaches, constructing gene expression profiles for several cancer types has previously turn out to be a promising means for cancer classification.

Feng Chu et al., [13] used the fuzzy neural network to categorize the microarray data sets. Also select some specific genes from collection of genes by using t-test-based feature collection approach. FNN is used in lymphoma data set, small round blue cell tumor data set and liver cancer data set. The results of these three data sets prove that FNN can achieve 100% accuracy with lesser amount of genes.

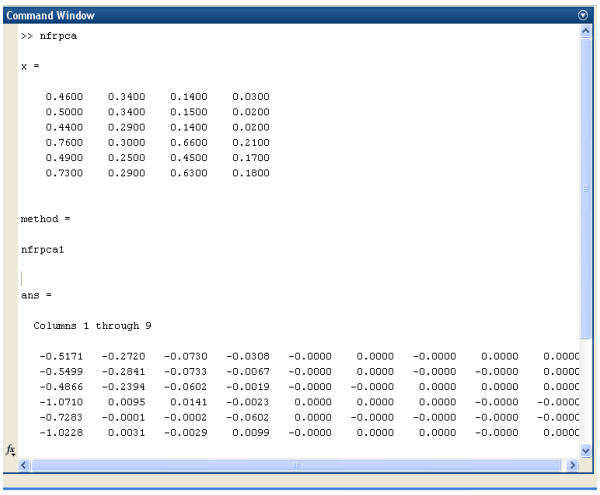

Figure 1. Dataset

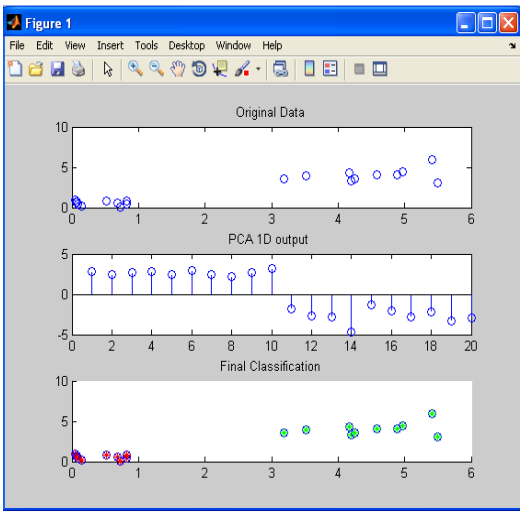

Figure 2. PCA Based Classification

4.2 Principal Component Analysis

Principal component analysis (PCA) [4] is an essential method in the framework of the various analysis methods. It is successfully used in many areas such as pattern recognition, process monitoring, data compression, and feature extraction, image processing and signal analysis. It is due to reason of its easiness and capability in processing massive amount of process data, PCA is identified as a dominant tool of statistical process care and widely used in the area for fault detection and diagnosis.The Proposed method acts well discriminating influence in gene expression research. The PCA-FNN [5] offers better classification accuracy than any other classifier.

5. Experimental Results

The experimental results are presented to establish the contribution of each factor used to optimize the FNN by using PCA method. First, we already quoted the reason for using the PCA for the model selection of a FNN. The comparison between GA search and PCA-FNN based model selection method in terms of average CPU time and testing accuracy. It has been empirical that many permutations may offer with similar validation and accuracy and the final sample can be trained by any one among them. The results in all the data sets prove that the FNN-PCA can achieve 95.8% accuracy with a much lesser amount of genes. The experimental study shows that the PCA-FNN model is very effective in terms of both evaluation time and classification performance.

6. Conclusion

Cancer research is one of the major research area in the medical science. Accurate prediction of various tumor types has maximum value in providing extended treatment and harmfulness reduction on the patients. In the earlier, cancer categorization is generally depends on Morphological and clinical analysis.

In this paper a new approach for process monitoring was proposed. This approach was a kind of Principal component analysis based on FNN. The reason for using fuzzy system was the Power of this system in approximating nonlinearity with arbitrary accuracy. We proposed PCA-FNN method to optimize the parameters. The PCA-FNN is proposed to categorize the liver cancer and through the analysis, the improved method arises the recognition rate in some dimension. In this paper, only binary classification problems were considered for the experiments but multiclass problems will be investigated in the future.

We conclude that the PCA-FNN classifier not only helps biological researchers differentiate cancers that are difficult to be classified using traditional clinical methods, but also helps biological researchers focus on a small number of important genes to find the relationships between those important genes and the development of cancers.

7. Scope for Future Enhancement

As a result of the success of the research, more areas of investigation can be pursed. This could be in terms of improving when large numbers of similar quantities are being estimated that is highest and lowest effects tend to be too extreme. This research which focuses on the cancer classification based on the Fuzzy Neural Network and Principal Component Analysis classifiers. In order to improve the learning capacity and to reduce complexity, this research requires some future enhancement.

8. References

[1] C. Debouck and P.N. Goodfellow, DNA Microarrays in Drug Discovery and Development," Nature Genetics Supplement, vol. 21, pp. 48-50, 1999.

[2] L. Wang, F. Chu, and W. Xie, "Accurate Cancer Classification Using Expressions of Very Few Genes," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 4, no. 1, pp. 40-53, Jan.-Mar. 2007

[3] Saravanan, V.; Mallika, R.; "An Effective Classification Model for Cancer Diagnosis Using Micro Array Gene Expression Data", International Conference on Computer Engineering and Technology (ICCET '09), Vol. 1, Pp. 137 – 141, 2009.

[4] M.A.Kramer , (1991) "Nonlinear Principal Component Analysis using Autoassociative Neural Networks", AIChE Journal, Vol. 37, No. 2, PP. 233-243.

[5] T.Cundari,(2002), "Robust Fuzzy Principal Component Analysis (FPCA). A Comparative Study Concerning Interaction of Carbon –Hydrogen Bonds with Molybdenum-Oxo Bonds", J.Chem.Inf.Comput.Sci, Vol. 42.

[6] Bevilacqua, V.; Mastronardi, G.; Menolascina, F.; Pannarale, P.; Pedone, A.; "A Novel Multi-Objective Genetic Algorithm Approach to Artificial Neural Network Topology Optimisation: The Breast Cancer Classification Problem", International Joint Conference on Neural Networks (IJCNN '06), Pp. 1958 – 1965, 2006.

[7] Kermani, B.G.; White, M.W.; Nagle, H.T.; "Feature extraction by genetic algorithms for neural networks in breast cancer classification", IEEE 17th Annual Conference Engineering in Medicine and Biology Society, Vol. 1, Pp. 831 – 832, 1995.

[8] Lipo Wang; Feng Chu; Wei Xie; "Accurate Cancer Classification Using Expressions of Very Few Genes", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 4, No. 1, Pp. 40 – 53, 2007.

[9] P. Broberg, "Statistical methods for ranking differentially expressed genes," Genome Biology, Vol.4, No. 6, 2003.

[10] Feng Chu; Lipo Wang; "Applying RBF Neural Networks to Cancer Classification Based on Gene Expressions", International Joint Conference on Neural Networks (IJCNN '06), Pp. 1930 – 1934, 2006.

[11] A.O. Hero, "Gene selection and ranking with microarray data," Seventh International Symposium on Signal Processing and its Applications, Vol. 1, pp. 457 – 464, 2003.

[12] Liver cancer dataset (http: //genomewww.stanford. edu/hc/):

[13] Feng Chu; Wei Xie; Lipo Wang; "Gene selection and cancer classification using a fuzzy neural network", IEEE Annual Meeting of the Fuzzy Information Processing (NAFIPS '04), Vol. 2, Pp. 555 – 559, 2004.