ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Optimization of Human Liver Cancer using Fuzzy Inference System for Microarray Gene Data

[1]Senbagavalli, M.E., Ph.D. and [2]Dr.G.Tholkappia Arasu., Ph.D.

[1]Assistant Professor, Department of IT, Jayam College of Engineering and Technology, Dharmapuri, India.
[2]Principal, AVS Engineering College, Salem, India.

ARTICLE INFO

Article history:
Received 28 January 2015
Accepted 25 February 2015
Available online 6 March 2015

Keywords:
Microarray Dataset, T-Score, Enrichment Score, Correlation Based Ranking, Support Vector Machine, Fuzzy Neural Network

ABSTRACT

In the fast growing world, cancer is one of the threatening diseases. Even though much medical advancements were discovered, it leads a great challenge for the human being. The uncontrolled and irregular growth of cells in tissue causes a cancer which leads to death of human being. Of all kinds of cancers, liver cancer is still to be considered and many researches are going to classify and predict its propagation and growth. Many classifying and ranking techniques like T-Score, ANOVA and Enrichment score were already deals with the optimization of cancer diseases. But these techniques are not to the mark of accuracy and classification. The classifying technique called Fuzzy System considers various parameters which are the major factors of cancer. Also it optimizes the causes and classifies the cancer cells. The experiment shows that the proposed technique classifies and optimizes the cancer cells significantly when compared to the conventional methods of cancer causing cells classification. It is considered that the dataset of DNA microarray gene data has more attraction in medical and scientific fields.

To Cite This Article: Senbagavalli, M.E., Ph.D. and Dr.G.Tholkappia Arasu., Ph.D., Optimization of Human Liver Cancer using Fuzzy Inference System for Microarray Gene Data. ***Aust. J. Basic & Appl. Sci.,*** *9(10): 86-92, 2015*

INTRODUCTION

Cancer plays a challenging role in medical science. So far many researches and innovation were introduced to tackle the cancer. But these innovations and results couldn't give complete solution for the cancer. The better way to reduce cancer death is to detect it at early stage and taking necessary remedies. The lab researches and clinically produced results are not sufficient to give best solution for cancer. The add-ons like statistical researches and computational techniques are become a great support to the clinical researches.

The recent survey says the maximum natural death in India is due to tobacco and alcohol consumption. The main factor for liver cancer is uncontrolled use of drugs and alcoholic contents. Predicting the outcome of a disease is one of the most challenging and interesting tasks where data mining techniques have to be applied.

Early diagnosis of liver cancer needs an accurate and reliable diagnosis solution that can be used by physicians to distinguish benign tumors from malignant ones without going for surgical biopsy. The objective of these predictions is to put affected patients to one of the two group likely *"benign*" *"malignant*". The patients grouped under benign are called noncancerous and the people group under malignant are cancerous (Fortina, P., 2002).

Many researches had undergone the utilization of computational tools, simulations and large value of medical data from various resources. As a result of this, data mining techniques has become a popular research tool for medical researchers to identify and exploit patterns and relationships among large number of variables, and made them able to predict the outcome of a disease using the historical datasets (Keller, J.M., 1985).

They computerize the diagnostic process and hence improve the accuracy and precision of conventional diagnostic methods. They facilitate thousands of gene expressions which are collected from the existing models (Ahmad M. Sarhan, 2009; Ying Xu, 2004). The added advantage of such microarray technology is the ability to classify the cancer types and cancer causing cells using the micro array gene expression datasets, which ultimately improve the diagnostic measures and yields very good results. The recent researches have proposed the purpose of classifying the cancer types using

Corresponding Author: Senbagavalli M.E., Ph.D., Assistant Professor, Department of IT, Jayam College of Engineering and Technology, Dharmapuri, India.
E-mail: senba1983@gmail.com

gene expression datasets (Yendrapalli, 2007; Sandrine Dudoit, 2002; Peterson and Ringner, 2003). This work intends to extend the work by performing a comparative analysis between the proposed classifying approach based dimensionality reduction and the conventional ranking techniques like T-Score, Enrichment Score and Support Vector Machines (SVM).

The objective of this study is to classify the cancer causing factors and optimizing the cancer cells. In this paper , An overview of the current research being carried out on various liver cancer datasets using the data mining techniques to enhance the liver cancer diagnosis and prognosis.

A. Challenges in cancer classification:

So far many researches were made on classification of cancer and its impact on human's health. All these researches were done on the classification problem by the statistical, machine learning and database research. However, the research on gene classification has new challenges due to its unique problem nature. Some of the challenges are summarized below:

The first challenge is the unique nature of the available gene expression data set is the foremost challenge (Yen, G. and T. Poggio, 2001). Second is the huge number of attributes (genes) which are irrelevant. Third challenge arises from the application domain of cancer classification. Though Accuracy plays a vital factor in cancer classification, the biological relevancy is another key criterion, as any biological information exposed during the process can help in added gene function discovery and other biological examinations.

Micro array data analysis has been effectively applied in a number of investigations over a wide range of biological disciplines, which comprises of cancer classification by class detection and prediction, recognition of the unknown effects of a specific therapy, recognition of genes suitable to a certain diagnosis or therapy, and cancer diagnosis. Several algorithms have been established for recovering data because it is expensive and time consuming to repeat the experiment (Statnikov, C., 2005). In this research, efficient neural network techniques are used with effective learning algorithms for providing significant cancer classification.

B. Data mining in cancer classification:

Han (Meyer, T. and I.R. Hart, 1998) called it "the extraction of interesting information or patterns from data in large databases". There are various other definitions for data mining from various researchers while Hand, Mannila and Smyth (2001) defined Data mining as "the analysis of observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner".

Bioinformatics is the application of molecular biology, artificial intelligence, computational techniques, statistics and mathematics to model, organize, understand and identify interesting knowledge associated with large-scale molecular biology databases called microarray data. In Bioinformatics classification is a major part and thus classification techniques play a vital role in bioinformatics, often using similarities of structure to infer similarity of function. A wide range of such techniques are used, both deterministic and probabilistic (Rui Xu, 2007).

Data mining in bioinformatics is susceptible due to many facets of biological databases which includes their size, their number, their diversity and the lack of a standard ontology to aid the querying of them, as well as the heterogeneous data of the quality and provenance information they contain.

Usually the DNA microarray data will contain complex data and unstructured data. It requires statistical problems. The microarray data is varied from the conventional biomedical data. The analysis of microarray data is complex in case of dimensionality Many analysis techniques treat each sample as a single point in a space with thousands of scope, then attempt by various methods to reduce the dimensionality of the data to something humans can visualize.

Related Work:

The recently proposed researches consider the substantial work and comparative analysis of difference set of microarray gene. This yielded inaccurate result. For the research purpose these methods considers only the statistical and biological techniques. This results in poor performance. While classifying the gene data type it is more important to have accuracy and reliability rather than time complexities and space complexities. The generation of the Fuzzy rules is very tedious 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. The new method has been proposed by combining kNN and SVM classifiers to improve the accuracy of classification.

Proposed System:

Our proposed system begins with the classification of micro array gene expression data. 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. 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.

While classifying the gene data type it is more important to have accuracy and reliability rather than time complexities and space complexities. The new method has been proposed by combining Support Vector Machine (SVM) and k-Nearest Neighbor (kNN) and classifiers to improve the accuracy of classification (Barati, E., 2011).

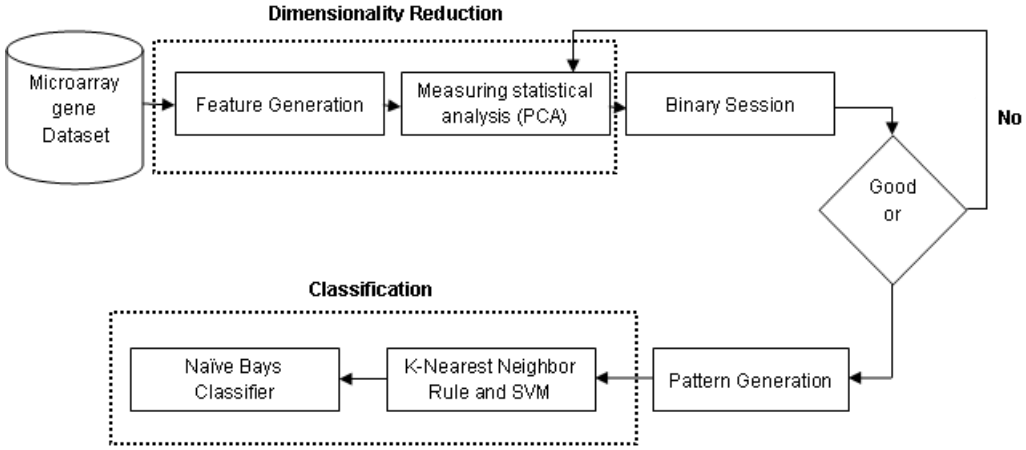

Fig. 1: Proposed System Architecture.

B. 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". The experiment is conducted with microarray gene data. Data mining techniques were implied to smoothen the collected data set. Three nonlinear smooth support vector machines (SSVMs) is used for classifying liver cancer patients into the three prognostic groups i e Good, Poor and Intermediate. These results suggest that the patients in the Good group should not receive chemotherapy while those in the Intermediate group should receive Combination has been performed. The classifier is trained using all possible gene combinations and the classifier is validated using 5 fold or 10 fold cross validation methods. The best gene combination can be selected from the result of accuracy (Jin Oh Kang, 2009).

From the various source of information it is understood that Machine learning based algorithms for gene selection and cancer classification as K-nearest neighbors, neural network, nearest shrunken centroids,

From this it is concluded that SVM with Fuzzy Inference System is the most often used classifier on better gene selection and higher classification accuracy.

A. K-nearest neighbor rule and SVM:

The k-NN rule is used to introduce notation. In the k-NN rule the pattern which is going to classify are symbolized as vectors in a d-dimensional Euclidean space Rd. chemotherapy based on our survival curve analysis.

Collective methods, the relevance value of gene is dependent on all other genes.

Datasets is randomly divided into two, one for training and another part for testing and gene ranking that is ANOVA P-Values can be computed using one-way ANOVA. Top genes were selected from the ranked data and gene logistic regression, and support vector machine (SVM) were proposed.

Methodology:

The methodologies used in this thesis for cancer classification is presented in this chapter. The process is divided in two phases. In the first phase, gene classification and ranking is carried out with the help of classification ranking technique such as k-means, Enrichment Score, Analysis Of Variance (ANOVA) and Correlation. In the second phase, classification is performed using Support Vector Machine (SVM) along with the class separability for classifying the cancer tumor exactly by means of microarray gene expression which be performed better than the usage of Fuzzy Neural Network (FNN) technique. In this proposed approach, the Support Vector Machine (SVM) classifier is used to test n-feature combinations.

A. Reduction methods for machine learning:

The initial data which is collected from various sources are undergone to the various data preprocessing techniques. Hence input space reduction is important in the building of accurate classifier. The collected datasets may contain several thousands of genes. Out of which many gene data are irrelevant. Therefore excluding the irrelevant genes will improve the performance of the classifier.

The dimension reduction methods can be categorized using as follows:

- *Feature selection or feature extraction:* Feature selection selects a subset of best original features.
- *Wrapper methods or Filter methods:* In wrapper methods, the reduction dimension is classifier dependent.
- *Individual methods or collective methods:* In individual methods, the relevance value of a gene is independent of all other genes. In

Several reduction dimension methods have been presented in the literature. This paper considers an individual feature extraction method to improve prediction accuracy.

B. Gene class identification:

The very first step is regrouping of similar genes in classes. The intention of grouping the similar gene is that genes which belong to the same class contain partially redundant information. In order to clustering the genes based on the similarity k-means algorithm is used. Once the clustering is completed, the next step is creating prototype from this cluster. This prototype can be expressed as

$$P_u = \left(y_{1,}\ y_{2,\ldots..,}\ y_M\right) \text{ with } y_j = \frac{1}{\text{Size }(C_u)} \sum_{\text{gene}(i)\in C_u}^{n} x_{i,j} \quad (1)$$

1. CM← Select the clustering method (k-means, NB Classifier
2. Naïve Bayes Clustering ← select the desired no. of cluster
3. For each iteration of the cross validation
   3.1. Define train set and test set
   3.2. Do Naïve Bayes Cluster of genes on train set using CM
   3.3. For each cluster C_u
      3.3.1. Build prototype P_u ←mean of this cluster
   3.4. Model ← training of SVM using prototype
   3.5. Accuracy ← prediction on the test set
4. Compute the average accuracy

C. Ranking of clustered genes:

1) Correlation:

A most commonly used measure of ranking among the data is Pearson's product Moment Correlation Coefficient (PMCC). This is denoted by *r* and calculated from sample data using the formula

$$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} \quad (2)$$

where,

$$s_{xx} = \sum (x_i - \bar{x})^2, \quad s_{yy} = \sum (y_i - \bar{y})^2 \quad (3)$$

$$s_{xy} = \sum (x_i - \bar{x})((y_i - \bar{y}) \quad (4)$$

2) Support Vector Machine:

SVM is one of the best linear classification methods. The transformation of the samples space to high-dimension space is possible by the kernel mapping and the best linear classification surface of samples in this new space is obtained. This Non-linear transformation is achieved by suitable inner product function. The best linear classification surface function of characteristics space can be described by the following equation:

$$g(x) = \sum_{j=1}^{n} a_j y_i k(x, x_i) + b \quad (5)$$

Where (x_i , y_i) are the two types of sample collection divided in the sample space, b is the classification threshold, and $k(x, x_i)$ is being the nonlinear kernel function that replace characteristics space and meet Mercer conditions.

Experiments:

This chapter provides the evaluation result for the proposed approach. The dataset used for evaluation is Liver Cancer Dataset. Initially, Correlation technique is used for ranking the important genes. Next, the classifier called Support Vector Machine (SVM) is used for classifying the occurrence of cancer.

An evaluation study of the proposed approach is presented in this chapter. The results of an extensive set of simulation tests are shown, in which the weather prediction approaches are compared under a wide variety of different scenarios.

A. Dataset:

The datasets that used in this study described in Table 1. All datasets were produced by oligonucleotide based microarray technology. Here, 6 datasets which has 2-9 distinct categories, 50-102 samples and 2308-10509 genes. All datasets are downloaded from http://www.gems-system.org.

B. Measurement of Accuracy:

For all genes in datasets, it is applied k-nearest neighbor and fuzzy k-nearest neighbor algorithm and it is determined each gene's category by using the other genes in the dataset as a training sample. Then it is calculated the prediction accuracy according to the relationship (2) for these two classification methods.

$$\text{Accuracy} = \frac{\sum_{s=1}^{N} p(s)}{N} * 100 \quad (6)$$

Where, p(s) value is assigned 1 if s-th gene's category (class) determined correctly, otherwise it is assigned 0. And N is the total number of genes in a dataset.

C. Evaluation:

The liver cancer data set has two classes, i.e., the non-tumor liver and HCC. The data set contains 156 samples and the expression data of 1,648 important genes. Among them, 82 are HCCs and the other 74 are non-tumor livers. The data is randomly divided into 78 training samples and 78 testing samples. In this data set, there are some missing values. K-nearest neighbor method is used to fill those missing values. The performance of the proposed approach is evaluated against the conventional techniques enrichment.

Table I: Enrichment Technique.

No. of fold	No. of Genes	Gene Combination	FNN	FIS
6	6	3	100	66.3
6	9	2	100	65

Table II: Correlation Technique.

No. of fold	No. of Genes	Gene Combination	FNN	FIS
6	6	3	93.42	100
6	9	2	95.45	100

Result and Analysis:

The experiment was conducted on the collected microarray gene data and following results were produced.

A. Accuracy:

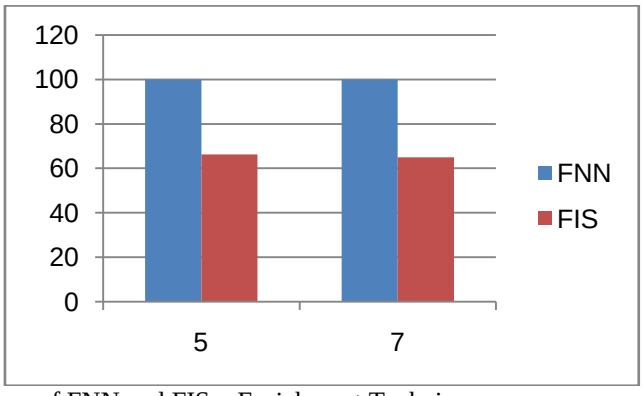

Fig. 2: Accuracy comparison of FNN and FIS – Enrichment Technique.

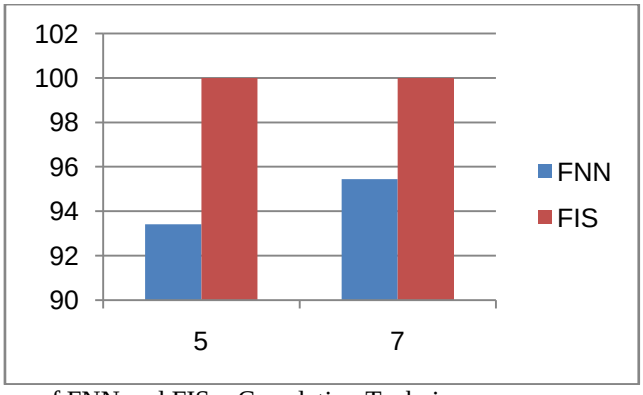

Fig. 3: Accuracy comparison of FNN and FIS – Correlation Technique.

B. Learning Time:

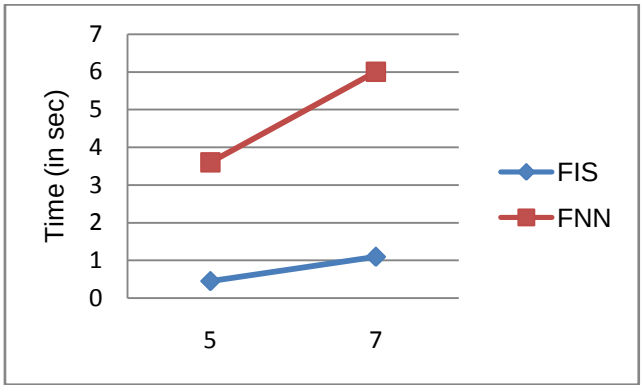

Fig. 4: Learning time comparison of FNN and FIS – Enrichment Technique.

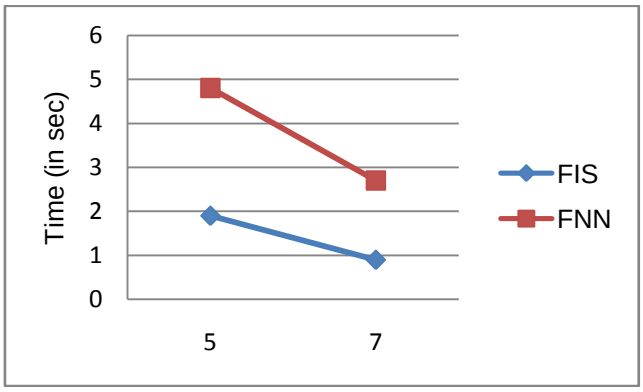

Fig. 5: Learning time comparison of FNN and FIS – Correlation Technique.

Conclusion:

From the techniques mentioned above, it is considered that the micro array data has been classified and clustered efficiently by the method of structural similarity. The outcome of this paper shows the better identification and classification of microarray data. Also it ensures the diagnosis by analysis and optimizing the kinds of cancer causing genes. The proposed Fuzzy Inference System along with ranking techniques like Enrichment and Correlation techniques produced more accurate result better than accuracy produced just by Support Vector Machine and Fuzzy Neural Network System.

REFERENCES

Ahmad M. Sarhan, 2009. "Cancer classification based on microarray gene expression data using dct and ann", Journal of Theoretical and Applied Information Technology, 6(2): 207-216.

Allan Tucker, Veronica Vinciotti, Xiaohui Liu, David Garway-Heath, 2005. A spatio-temporal Bayesian network classifier for understanding visual field deterioration, Artificial Intelligence in Medicine, 34(2).

Barati, E., 2011. "A Survey on Utilization of Data Mining Approaches for Dermatological (Skin) Diseases Prediction" Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Health Informatics (JSHI): March Edition.

Dr. Medhat Mohamed Ahmed Abdelaal, " Using data mining for assessing diagnosis of breast cancer" Proceedings of International Multi Conference on Computer Science and Information Technology, pp: 11-17, ISBN 978-83-60810-27-9 ISSN 1896-7094.

Dursun Delen, 2009. Analysis of cancer data: a data mining approach, The Journal of Knowledge Engineering, Expert Systems, February, 26(1).

Fortina, P., S. Surrey and L.J. Kricka, 2002. "Molecular diagnostics: hurdles for clinical implementation", *Trends Molecular Medicine*, 8: 264-266.

Hand, David, Heikki Mannila and Padhraic Smyth, 2001. Principles of Data Mining, MIT Press.

Jin Oh Kang, 2009. "Prediction of Hospital Charges for the Cancer Patients with Data Mining Techniques", 15(1): 13-23.

Keller, J.M., M.R. Gray and J.A. Givens, 1985. "A fuzzy k-nearest neighbor algorithm", *IEEE Transactions on Systems, Man and Cybernetics*, 15: 580-585.

Meyer, T. and I.R. Hart, 1998. “Mechanisms of Tumour Metastasis”, European Journal of Cancer, 34: 214-221.

Peterson and Ringner, 2003. “Analyzing Tumor Gene Expression Profiles”, Artificial Intelligence in Medicine, 28(1): 59-74.

Remco R. Bouckaert, 2010. “Weka-Experiences with a Java Open-Source Project” Journal of Machine Learning Research, 11: 2533-2541.

Rui Xu, Anagnostopoulos, G.C. Wunsch, D.C.I.I., 2007. “Multiclass Cancer Classification Using Semisupervised Ellipsoid ARTMAP and Particle Swarm Optimization with Gene Expression Data”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4(1): 65-77.

Sandrine Dudoit, Jane Fridlyand and Terence P. Speed, 2002. “Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data”, Journal of the American Statistical Association, 97: 77-87.

Statnikov, C., F. Aliferis, I. Tsamardinos, D. Hardin and S. Levy, 2005. “A comprehensive evaluation of multilevel category classifier techniques for microarray gene expression cancer diagnosis”, Bioinfo, 21: 631-643.

Thair Nu Phyu, 2009. “Survey of Classification Techniques in Data Mining” Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, Hong Kong.

Yen, G. and T. Poggio, 2001. “Multiclass classification of SRBCT tumors”, *Technical Report Al Memo 2001-018 CBCL Memo 206*, MIT Press.

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

Ying Xu, Victor Olman and Dong Xu, 2001. "Minimum Spanning Trees for Gene Expression Data Clustering", Genome Informatics, 12: 24-33.