IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

A Survey on fuzzy expert system for improving microarray data classification accuracy

Deepakkumar.S[1], Mohankumar.M[2]

PG Scholar, Department of Computer Science and Engineering, SVCET, Virudhunagar, India[1]

Assistant Professor, Department of Computer Science and Engineering, SVCET, Virudhunagar, India [2]

Abstract: Building an accurate fuzzy expert system will improvise the classification of microarray data and reduces the complexity. Modern practice in the classification of microarrays' data has two main limitations: (1) the dependability of the training data sets for building classifiers, (2) the model to be classified does not fit in to any of the existing classes. Medical thermography is very useful in a variety of medical applications as well as the detection of breast cancer by identifying the local temperature and the elevated metabolic commotion of cancer cells. Distinct conventional expert systems, which are mainly symbolic reasoning engines, fuzzy expert systems are oriented toward numerical processing. To address the interpretability-accuracy trade-off, the system proposes hybrid Ant Bee Algorithm (ABA) and it is evaluated using six gene expression data sets.

Keywords: Microarray data, fuzzy expert system, ant colony optimization, artificial bee colony, mutual information.

I. INTRODUCTION

Microarray technology generates an enormous quantity of data by measuring, all the way through the hybridization process, the levels of nearly all the genes expressed in a biological model. One can expect that knowledge gleaned from microarray data will put in considerable advancement in elemental questions in biology. One significant goal of analyzing microarray data is to classify the samples. The terminology developed in these papers consists in discrimination methods and machine learning methods. In microarray studies, the number of samples, *n*, is relatively small compared to the number of genes, *p*, usually in thousands. Unless a preliminary variable selection step is performed, standard statistical methods in classification perform poorly because there are far more variables than observations. One problem is multicol linearity: estimating equations become singular and have no unique and stable solution. For instance, the pooled within-class sample covariance matrix in Fisher's linear discriminate function is singular if $n<p + 2$. Even if all genes can be used as in support vector machines, it seems to be not sensible to use all the genes.

In fact, this allows the presence of the noise associated with genes of no discrimination power. That inhibits and degrades the performances of the classification rules in their application to unclassified tumour. Dimension reduction is desired to reduce the high *p*-dimensional gene area. In the past mentioned works, the authors have used uni-variate methods for reducing the number of genes. Alternative approaches to handle the dimension reduction problem can also be used.

The method of partial least squares have been found to be a useful dimension reduction system. The principal component regression is for statistical view of partial least squares and principal component regression. The function of principal component regression is to make orthogonal tumour descriptors that reduce the dimension to only a few gene components (super-genes). However, dimension reduction is achieved without regard to the response variable and may be disorganized. This is the motivation, the partial least squares more adapted than principal component regression for dimension reduction based prediction. Indeed, partial least squares components are chosen so that the sample covariance between the response and a linear combination of the *p* predictors is maximum.

The tiny, round blue cell tumours of childhood, which include non-Hodgkin lymphoma (NHL) and Ewing family of tumours (EWS), are consequently named because of their related appearance on routine histology˙ However, exact diagnosis of SRBCTs is important because the treatment options, responses to therapy and prognoses vary widely depending on the diagnosis. As their name implies, these cancers are hard to differentiate by light microscopy, and currently no single test can correctly differentiate these cancers. RT-PRINCIPAL COMPONENT REGRESSION are worn regularly more for diagnostic affirmation following the detection of tumour-specific translocations in alveolar rhabdomyosarcoma (ARMS).

Still, molecular markers do not forever offer a perfect diagnosis, as on circumstance there is failure to detect the usual translocations, due to either technical difficulties or the occurrence of variant translocations. Gene expression profiling using microarrays allows a simultaneous analysis of multiple markers, and has been used to classify cancers into subgroups. However, in spite of the many statistical techniques to examine gene-expression data, not any has been thoroughly tested for their capacity to accurately differentiate cancers belong to numerous diagnostic categories.

DNA microarrays are one of the most important technologies for genetic study. These are small solid supports, on which sequences of DNA are fixed in an orderly manner. Millions of DNA probes can be fond of a single slide and used to examine and calculate the activity of genes. Scientists are using DNA microarrays to examine several phenomena (e.g., cancer, pest control, tumour etc.) by measuring changes in gene expression. Then by learning how cells react to a disease or to a picky treatment.

Classification microarray data leads to quite a lot of challenges to usual machine learning methods. In exacting, microarray classification faces the small N, large P (NP) problem of statistical learning. The system proposes hybrid Ant Bee Algorithm (ABA) and it is evaluated using six gene expression data sets in this project.

II. RELATED WORK

Di Carlo S, et.al (2011) describes methodology for generating data classifier for clinical diagnostics. Although great advances in finding out cancer molecular profiles, the proper function of microarray technology is still a very big challenge. They suggest some effective methodologies to overcome the limitations of Current practices in the classification of microarrays data. To address the problems of state-of-the-art classification algorithms, this paper presents classification algorithm based on graph theory. Using of log-ratio seems to be a better solution to discover relevant genes. Decision rule is used to imitate a human cognitive process to carry out predictions. It can also provide important diagnostic information such as the identification of genetic similarities with known diseases. The limitation is it needs to know the Boolean cut-off among relevant and not relevant genes.

S Lambert-Lacroix, et.al (2005) proposed a statistical dimension-reduction approach for the classification of tumours. Authors view the classification problem as a regression one with few observations and many predictor variables. They proposed a new method combining partial least squares and Ridge penalized logistic regression. The idea developed here is to penalize with a Ridge penalty the likelihood criterion in order to constrain the pseudo-response variable to be finite. This overcomes the problem of a high-dimensional gene expression space so common in such type of problems. It allows the combination of a regularization step and of a dimension-reduction step.

Frank Westermann, et.al (2011) proposed this study to optimize the classification of cancers. They used stringent quality filter to comprise only the genes for which there were good measurements for all samples. This may remove certain genes that are highly expressed in some cancers, but not expressed in other cancers. They developed a method of diagnostic classification of cancers from their gene expression signatures. Although it achieved high sensitivity and specificity for diagnostic classification and believed that with larger arrays and more samples it will be possible to improve on the sensitivity of these models for purposes of diagnosis in clinical practice. The only limitation is it does not always provide a definitive diagnosis.

Y Wang, et.al (2005) proposed an approach to overcome curse-of-dimensionality problem. By using this selection techniques feature we can select a small subset of genes for classification. To achieve this goal, they developed a novel hybrid approach that combines gene ranking and clustering analysis. They applied filtering algorithms to select a set of top-ranked genes, and then applied hierarchical clustering on these genes to generate a dendrogram. In this approach, first feature filtering algorithms are applied to select a set of top-ranked informative genes. This approach was capable of selecting very small sets of marker genes without sacrificing classification accuracy. The problem is, to prohibitively expensive to try all possible numbers for clusters in order to find a setting that provides the best classification performance.

D. E. Johnson F. J. Oles, et.al (2011) proposed a method for rule induction based on decision-tree. They said it may automatically categorize text documents. The rule induction involves the new combination of (1) a fast decision tree induction algorithm especially suitable to text data and (2) an innovative method for converting a decision tree to a rule set that is equivalent to the original tree. Each rule ultimately produced by such a system states that a condition, which is usually a conjunction of simpler conditions, implies membership in a particular category. The final criterion for evaluating a tree is by its classification performance. The modified entropy is introduced to balance the advantage of classification error. It invents an industrial strength state-of-the-art prototype system that eventually evolved into the IBM Text Analyser product offering. It is a fast decision tree induction algorithm especially suited to text data. The classification error function does not favour a partition that enhances the purity.

DQ Naiman, et.al (2011) compared some other machine learning techniques for class prediction in 19 binary and multi-class gene expression datasets involving human cancers. These studies have shown that cancer tissue samples can be effectively detected and distingue by their gene expression patterns via machine learning approaches. Here they introduced a new classifier in order to address these problems, *k*–Top Scoring Pairs. The *k*–Top Scoring Pairs classifier performs as efficiently as Prediction Analysis of Microarray and support vector machine, and outperforms other learning methods such as decision trees, *k*-nearest neighbour and naïve bayes. It provides decision rules that usually involve many fewer genes and are far easier to interpret. This classifier is a very useful tool for cancer classification from microarray gene expression data. The limitation is it may be confused with rank-based methods.

Youngmi Yoon, et.al (2010) proposed new methodologies to classify microarray data based on k-Top-Scoring rank comparison decision rules. The Proposed phenotype classifier is an ensemble method with k-top-scoring decision rules. Each rule involves a number of genes, a rank comparison relation among them, and a class label. Generalizing the number of genes increases the robustness and reliability of the classifier. The classifier consists of k-decision rules.

DOI 10.17148/IJARCCE.2015.41141

The proposed method does not generate all of the possible rules of gene combination. It only builds a new longer rule by combining shorter rules while estimating the score range of the new rule. This reduces computing time and memory space. This paper, generalize the number of genes involved in each rule. It is reliable methodology. Sometimes it may not generate all the likely rules of gene combination.

Staal. A, et.al (2004) presented a fuzzy set framework for the implementation of classifiers. Their study has important limitations, that are the sample sizes were extremely small, and there may be an optimistic bias reflected in the results. Interpretation of classification models derived from gene-expression data is usually not simple, yet it is an important aspect in the analytical process. The performance of small rule-based classifiers based on fuzzy logic in five datasets that are different in size, laboratory origin and biomedical domain. The latter have theoretical limitations that, although difficult to verify in practice, may come into play in one or more datasets that we utilized. If, for example, the data are not linearly separable, then more flexible models such as logistic regression models with interaction terms, artificial neural networks, or support vector machines with non-linear kernels might be more appropriate. One of the main drawbacks of this approach is the large number of rules it results in. In order to reduce the number of rules, one can filter them manually.

Schaefer, et.al (2009) proposed a computational approach to the diagnosis breast cancer based on medical infrared imaging. Asymmetry analysis of breast thermograms is performed using a wide variety of statistical features. These features are then fed into a fuzzy if-then rule based classification system which outputs a diagnostic prediction of the investigated patient. Thermography is very useful in various medical applications including the detection of breast cancer where it is able to identify the local temperature increase caused by the high metabolic activity of cancer cells. In this paper concentrates breast cancer analysis based on thermography, using a series of statistical features extracted from the thermograms quantifying the bilateral differences between left and right breast areas, coupled with a fuzzy rule-based classification system for diagnosis. Clearly the simplest feature to describe a temperature distribution such as those encountered in thermograms is to calculate its statistical mean. The symmetry features calculates the mean for both breasts and use the absolute value of the difference of the two. Similarly it calculates the standard temperature deviation and use the absolute difference as a feature. It is useful in various medical applications. It is not robust enough yet to account for the variety of cases.

Ganesh Kumar.P, et.al (2012) proposed a technique to accurately classify the sample microarray data using fuzzy system and genetic swarm algorithm. This paper proposes a new Genetic Swarm Algorithm for getting most favourable rule set and membership function. Advanced and problem specific genetic operators are proposed to improve the convergence of Genetic Swarm Algorithm and classification accuracy. The fuzzy system can produce interpretable classifier with knowledge expressed in terms of if-then rules and membership function. This combines the strength of both Genetic Swarm Algorithm and PSO. GSA uses a mixed form of representation for training the solution variables of the fuzzy expert system. It is used to generate the classifier model for gene expression data. It does not provide any measure on the deeper understanding of the fundamental questions in biology and medicine.

III. PROPOSED WORK

Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. But the scientists face few problems such as inaccurate classification and the gene expression data are very large in number. To overcome the problems we propose a fuzzy expert system with fuzzy type II if-then rules and some colony algorithms for optimization.

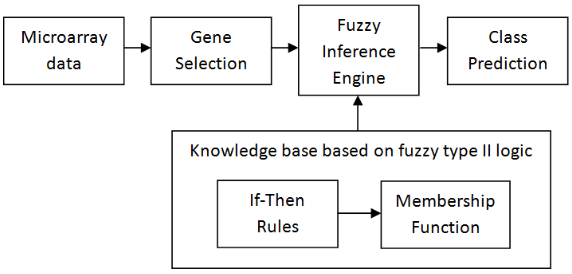

Fig. 1. Design of fuzzy expert system with type II logic

Gene selection has drawn particular concentration though designing a fuzzy expert system for Microarray data classification because of its high dimensional nature. Mutual information technique is for selecting revealing genes. A fuzzy expert system is an expert system that uses a collection of if-then rules and membership functions, instead of Boolean logic to reason about data. The general form of an if-then rule in the proposed fuzzy expert system is as below:

Rj: if xp1 is Aj1 and . . . and xpn is Ajn then class Cj where Aj1; . . . ;Ajn are antecedent fuzzy sets of the input genes xp1; . . . ; xpn and Cj is one of the output class label.

A collection of such rules forms the rule base for the fuzzy expert system upon which qualitative reasoning is performed to infer the results. The relation between input and output is expressed using a fuzzy relation constructed on the basis of fuzzy if-then rules. A fuzzy relation is a fuzzy set defined on universal sets, which are Cartesian products. Mathematically, a fuzzy set A in the universe of discourse X is defined to be a set of ordered pairs.

IV. CONCLUSION

Recent studies have shown that microarray gene expression data are useful for phenotype classification of many diseases. A major problem in this classification is that the number of features (genes) greatly exceeds the number of instances (tissue samples).The classification process requires maximum accuracy and minimum complexity (compact rules).Various methodologies and

DOI 10.17148/IJARCCE.2015.41141

classifications algorithms are tries to overcome these problems, but yet it is a complex task. To achieve maximum accuracy we need to build the system which is effective and limited rule set. For this we may use fuzzy logic and some colony algorithms. Colony algorithms help us to select best attributes (optimized) for the process.

ACKNOWLEDGMENT

My deepest gratitude is to my advisors, **Dr.P.Murugeswari.,Ph.D., Mr.M.Mohankumar**. I have been amazingly fortunate to have advisors who gave me the freedom to explore on my own and at the same time the guidance to recover when my steps faltered. They taught me how to question thoughts and express ideas. Their patience and support helped me overcome many crisis situations and finish this work.

REFERENCES

[1] T.L. Bergemann and L.P. Zhao, “Signal Quality Measurements for cDNA Microarray Data,” IEEE/ACM Trans. Computational Biologyand Bioinformatics, vol. 7, no. 2, pp. 299-308, Mar./Apr. 2010.

[2] A. Benso, S.D. Carlo, and G. Politano, “A cDNA Microarray GeneExpression Data Classifier for Clinical Diagnosis Based on GraphTheory,” IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 8, no. 3, pp. 577-591, May/June 2011.

[3] L. Li, “Gene Selection for Sample Classification Based on Gene Expression Data: Study of Sensitivity to Choice of Parameters ofthe ga/knn Method,” Bioinformatics, vol. 17, pp. 1131-1142, 2001**.**

[4] G. Fort and S.L. Lacroix, “Classification Using Partial Least Squares with Penalized Logistic Regression,” Bioinformatics,vol. 21, no. 7, pp. 1104-1111, 2005.

[5] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, and P.S. Meltzer, “Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks,”Nature Medicine, vol. 7, pp. 673-679, 2001.

[6] D.E. Johnson, F.J. Oles, T. Zhang, and T. Goetz, “A Decision-TreeBased Symbolic Rule Induction System for Text Categorization,”IBM Systems J., vol. 41, no. 3, pp. 1-10, 2002.

[7] Y. Yoon, S. Bien, and S. Park, “Microarray Data Classifier Consisting of k-Top-Scoring Rank-Comparison Decision Rules with aVariable Number of Genes,” IEEE Trans. Systems, Man, and Cybernetics-Part C: Applications and Rev., vol. 40, no. 2, pp. 216-226, Mar.2010.

[8] G. Schaefer, “Thermography Based Breast Cancer Analysis UsingStatistical Features and Fuzzy Classification,” Pattern Recognition,vol. 42, no. 6, pp. 1133-1137, 2009.

[9] P.GaneshKumar, T. Aruldoss Albert Victore, P. Renukadevi, andD. Devaraj, “Design of Fuzzy Expert System for Microarray DataClassification Using a Novel Genetic Swarm Algorithm,” ExpertSystems with Applications, vol. 39, no. 2, pp. 1811-1812, 2012.

[10] P. Maji, “f-Information Measures for Efficient Selection of Discriminative Genes from Microarray Data,” IEEE Trans. Biomedicine Eng., vol. 56, no. 4, pp. 1063-1069, Apr.2009

[11] D. Devaraj and B. Yegnanarayana, “Genetic Algorithm-Based Optimal Power Flow for Security Enhancement,” IEE Proc. Generation, Transmission and Distribution, vol. 152, no. 6, pp. 899-905, Nov.2005

[12] A. Sharma, S. Imoto, and S. Miyano, “A Top-r Feature SelectionAlgorithm for Microarray Gene Expression Data,” IEEE/ACMTrans. Computational Biology and Bioinformatics, vol. 9, no. 3,pp. 754-764, May/June 2012.

[13] P. Maji and S.K. Pal, “Fuzzy-Rough Sets for Information Measures and Selection of Relevant Genes from Microarray Data,” IEEETrans. Systems, Man and Cybernetics, vol. 40, no. 3, pp. 741-752,June 2010

DOI 10.17148/IJARCCE.2015.41141