Volume 6, Issue 7, July 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Evaluate Cancer Gene Expression using Fuzzy Rough Approach and Fuzzy TSVM

Krishnaveni Sakkarapani*
Assistant Professor
Department of Computer Applications
Pioneer College of Arts & Science
Coimbatore, Tamilnadu, India

Sathish Ayyaswamy
Assistant Professor
Department of Computer Science
Maharaja Arts and Science College
Coimbatore, Tamilnadu, India

Abstract— Cancer classification using gene expression data usually relies on traditional supervised learning techniques, in which only labeled data can be exploited for learning. They are also useful for identifying potential gene markers for each cancer subtype, which helps in successful diagnosis of particular cancer type existing system developed a classification system by identifying potential gene markers and subsequently applying the recent technique on the selected genes for the classification of human cancer. In this paper, we use a fast clustering based feature selection technique and Fuzzy based Transductive Support Vector Machine (FTSVM). By using the fast clustering based feature selection we can obtain the high probability of producing a subset of useful and independent features. To ensure the efficiency of FAST, we adopt the efficient Minimum-Spanning Tree (MST) clustering method. FTSVM method generates membership values iteratively based on the positions of training vectors relative to the TSVM decision surface itself. From the experiments we can obtain the high efficiency and effectiveness of this system. So the FTSVM has high accuracy compared to other system.

Keywords— SVM, TSVM, FTSVM, Rough set approach, Cancer gene

I. INTRODUCTION

Data mining is one of the most effective services available today. With the help of data mining, one can discover precious information about the customers and their behavior for a specific set of products and evaluate and analyze, store, mine and load data related to them. For its flexible nature as far as applicability is concerned is being used vehemently in applications to predict crucial data including industry analysis and consumer buying behaviors. Fast paced and prompt access to data along with economic processing techniques has made data mining one of the most suitable services that a company seeks.

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm and Nearest Neighbor method are used for knowledge discovery from databases. With the recent advancement of DNA microarray technologies, the expression levels of thousands of genes can be measured simultaneously. The obtained data are usually organized as a matrix, which consists of n columns and m rows. The columns represent genes and the rows correspond to the samples. Given this rich amount of gene expression data, the goal of microarray analysis is to extract hidden knowledge from this matrix.

The analysis of gene expression may identify mechanisms of gene regulation and interaction, which can be used to understand a function of a cell. Fig. 1 shows the general representation of cancer specific gene expression.

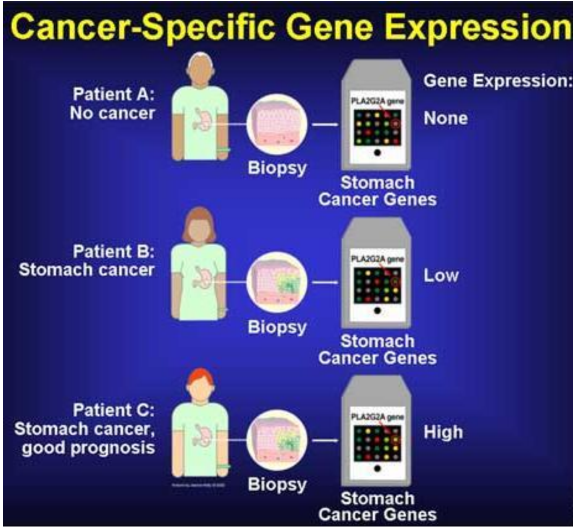

Fig. 1 Examples for Gene Expression

A microarray can contain up to 20,000 features, each of which recognizes mRNA from a single gene and relatively small number of experiments or samples. As a consequence, the identification of discriminates genes to classifying tissue types, e.g., presence of cancer is fundamental and practical interest. Such genetic markers, in fact can be found of value in further investigation of the disease and in future therapies[8].

II. LITERATURE SURVEY

Ability of measuring gene expression for a very large number of genes, covering the entire genome for some small organisms, raises the issue of characterizing cells in terms of gene expression that is, using gene expression to determine the fate and functions of the cells. The most fundamental of the characterization problem is that of identifying a set of genes and its expression patterns that either characterize a certain cell state or predict a certain cell state in the future[12]. Cancer classification using gene expression data usually relies on traditional supervised learning techniques, in which only labeled data can be exploited for learning, while unlabeled data are disregarded. Recent research in the area of cancer diagnosis suggests that unlabeled data, in addition to the small number of labeled data, can produce significant improvement in accuracy, a technique called semi supervised learning.

Major research on extending Support Vector Machines (SVMs) to handle semi-labeled data is based on the following idea to solve the standard Inductive SVM (ISVM) while treating the unknown labels as additional optimization variables. By maximizing the margin in the presence of unlabeled samples, one can learn the decision boundary that traverses through low density regions while respecting labels in the input space. A forward greedy reduction algorithm was exploited to identify the gene markers. The effectiveness of the FTSVM was compared with the LDS and ISVM on the basis of overall average accuracy. The Transductive SVM (TSVM) implements the cluster assumption more directly by trying to find a hyper plane which is far away from the unlabeled points. In our opinion, the rationale for maximizing the margin is very different for the labeled and unlabeled points: For the labeled points, it implements regularization. Intuitively, the large margin property makes the classification robust with respect to perturbations of the data points.

TSVM might seem to be the perfect semi-supervised algorithm, since it combines the powerful regularization of SVMs with a direct implementation of the cluster assumption. However, its main drawback is that the objective function is non-convex and thus difficult to minimize. Consequently, optimization heuristics like SVM light sometimes give bad results and are often criticized.

The main points of this paper are,

- The objective function of TSVM is appropriate, but different ways of optimizing it can lead to very different results. Thus, it is more accurate to criticize a given implementation of the TSVM rather than the objective function itself.
- The search for a low density decision boundary is difficult. The task of the TSVM algorithm can be eased by changing the data representation.

The advent of microarray technologies have now made it possible to have a global and simultaneous view of the expression levels for many thousands of genes over different time points during different biological processes. Clustering is a primary approach to analyze such large amount of data. Clustering is an unsupervised exploratory pattern classification technique which partitions the input space into the k Fuzzy C-Means (FCM) and its variants are widely used techniques used for microarray data clustering. In general, it has been observed that the performance of clustering algorithms degrade with more and more overlaps among clusters in a dataset[3].

Gene from DNA sequences is an important problem in bioinformatics. Human chromosomes have been sequenced; pegging the estimated number of genes to development of reliable automated techniques for interpreting long anonymous genomic sequences became imperative. Some other areas of immense interest within bioinformatics are the aforementioned level, include the optimization of drug administration schedule based on drug response data, automatic determination of cancer types from image data (e.g., skin cancer, breast cancer), and determining the stage of cancer from cellular images of the infected areas (e.g., cervical cancer)[5].

The Cancer-miRNA network is developed by mining the literature of experimentally verified cancer-miRNA relationships. This network throws up several new and interesting biological insights which were not evident in individual experiments, but become evident when studied in the global perspective. From the network a number of cancer-miRNA modules have been identified based on a computational approach to mine associations between cancer types and miRNAs. Besides this, neighboring miRNAs may also show a similar dysregulation patterns (differentially coexpressed) in the cancer tissues. Then in 67% of the cancer types have at least two neighboring miRNAs showing down regulation which is statistically significant (P < 10-7, Randomization test)[4].

According to the miRBase of Sanger Institute approximately 700 miRNAs are found in human and up to one third of the total human mRNAs are predicted to be miRNA targets. Each miRNA can target approximately 200 Transcripts directly or indirectly, whereas more than one miRNA can converge on a single protein coding gene target. Recent studies indicate that many miRNAs, referred to as onco/ tumor suppressor miRNAs, are involved in the development of various human malignancies([9]-[11],[13]).

The performance of the SiMM-TS (using VGAMOGA combination) is also compared with those of well known gene expression data clustering methods, namely, Average Linkage, SOM and a recently developed technique called Chinese Restaurant clustering (CRC) and also with VGA and IFCM applied independently. One artificial and three real life gene expression data sets are considered for experiments[14].

The goal of semi-supervised classification is to use unlabeled data to improve the generalization. The cluster assumption states that the decision boundary should not cross high density regions, but instead lie in low density regions. The virtually all successful semi-supervised algorithms utilize the cluster assumption, though most of the time indirectly. New classification systems are used to identify potential gene markers as well as subsequently applying the techniques on the selected genes for the classification of human cancer. A forward greedy reduction algorithm was exploited to identify the gene markers. The effectiveness of the technique compared with the LDS and ISVM on the basis of overall average accuracy. Main Contribution of this paper is explorations of new strategies and development of new methods to improve accuracy and effectiveness of algorithm results.

When comparing the algorithms like Inductive SVM, Consistency-Based Feature Selection, Transductive SVM, FAST Clustering Based Feature Selection and Fuzzy Transductive SVM (FTSVM). The time concern and accuracy are also evaluated. Hence, the performance criteria of the novel algorithm are also comparatively high. The main idea behind this work is to uncover the better quality algorithm to classify the Cancer Gene.

III. DATASET DESCRIPTION

For identifying the cancer we have used various datasets such as LEUKEMIA, SRBCT, MLL and DLBCL. Leukemias are primary disorder of bone marrow[6]. They are malignant neoplasms of hematopoietic stem cells. The total number of genes to be tested is 7129 and number of samples to be tested is 72, which are all acute leukemia patients, either Acute Lymphoblastic Leukemia(ALL) or Acute Myelogenous Leukemia(AML). Lymphoma is a broad term encompassing a variety of cancers of the lymphatic system. It is otherwise known as Small Round Blue Cell Tumours (SRBCT). Total number of genes to be tested is 4026 and the number of samples to be tested is 62. There are all together three types of lymphomas[1]. The first category, Chronic Lymphocytic Lymphoma(CLL) has 11 patients, the second type Follicular Lymphoma(FL) has 42 and the third type Diffuse Large B-cell Lymphoma(DLBCL) has 9. In Mixed Lineage Leukemia (MLL) dataset, total number of genes to be tested is 12582. The dataset contains 72 samples with 24 ALL, 20 MLL and 28 AML[2]. Lung cancer is a disease in which certain lung cells don't function right, divide very fast, and produce too much tissue forming a lung tumor otherwise called as DLBCL. There are 181 tissue samples among which 31 samples belong to MPM and 150 belong to ADCA. Each sample is described by 12533 genes[7]. The dataset information can be summarized in table I.

Table I. microarray dataset Description

Dataset	Samples	Genes	Classes	Description
LEUKEMIA	72	7129	2	47 ALL and 25 AML
SRBCT	62	4026	3	11 CLL, 42 FL, and 9 DLBCL
MLL	72	12582	3	24 ALL, 20 MLL and 28 AML
DLBCL	181	12533	2	31 MPM, 150 ADCA

Common symptoms of chronic or acute Leukemia showed detailed in Fig. 2.

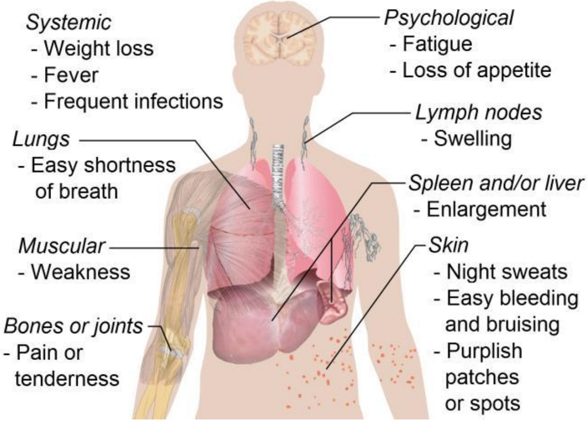

Fig. 2 Leukemia cancer

IV. PERFORMANCE ANALYSIS

All The performance of the algorithms has been evaluated in terms of sensitivity, specificity and accuracy. The three indices are defined as follows,

$$Sensitivity = \frac{TP}{TP+FN}$$

$$Specificity = TN/(TN + FP)$$

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

True Positive (TP)**:** The classification result is positive in the presence of the clinical abnormality
True Negative (TN)**:** The classification result is negative in the absence of the clinical abnormality
False Positive (FP)**:** The classification result is positive in the absence of the clinical abnormality
False Negative (FN)**:** the classification result is negative in the presence of the clinical abnormality

A. Sensitivity

Recall (sensitivity) value is calculated based on the retrieval of information at true positive prediction, false negative. In healthcare data precision is calculated the percentage of positive results returned that is recall in this context is also referred to the True Positive Rate. Recall is the fraction of relevant instances that are retrieved.

Fig. 2 shows the sensitivity rate of the systems such as ISVM, ISVM-CBFS, TSVM, TSVM-CBFS, Fuzzy, Fuzzy-fast and TSVM-fast based on two parameters of sensitivity and number of instances. In this figure, x axis will be number of instances and y axis will be sensitivity rate. When the instances are increased, the sensitivity rate of the system is decreased. From the graph we can see that, sensitivity of the system is reduced somewhat in Fuzzy based system, which will be the best one.

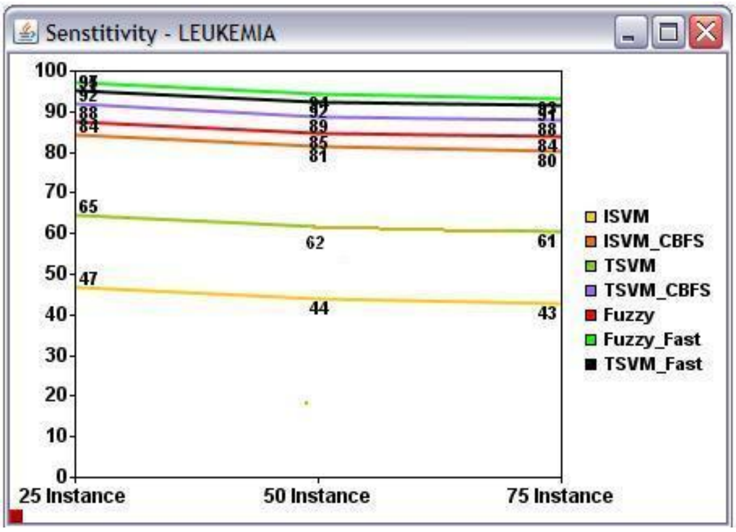

Fig. 2 Sensitivity Leukemia Chart

B. Specificity

Fig. 3 shows the specificity rate of the same systems which are used in sensitivity. The x axis will be number of instances and y axis will be Specificity rate. When the instances are increased, the specificity rate of the system is decreased. From the graph we can see that, Specificity of the system is reduced somewhat in other system than the Fuzzy based system. From this graph we can say that the Specificity of Fuzzy based system is increased which will be the best one.

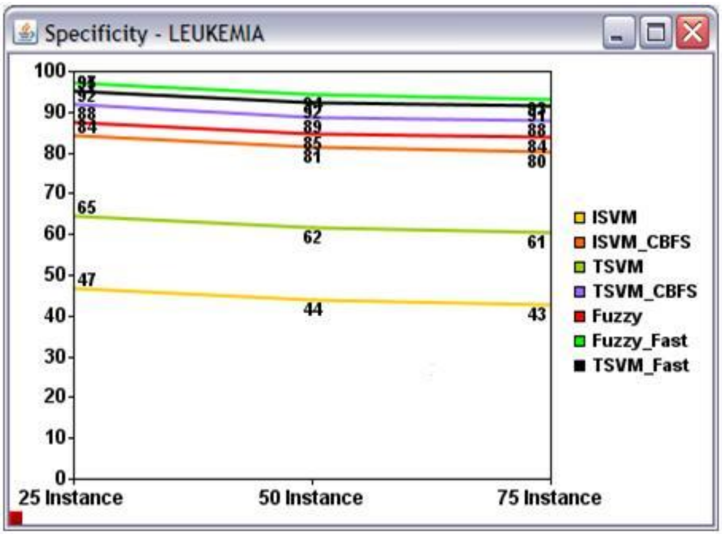

Fig. 3 Specificity Leukemia Chart

C. Accuracy

Fig. 4 shows x axis will be number of instances and y axis will be accuracy rate. When the instances are increased, the accuracy rate of the system is decreased. From the graph we can see that, accuracy of the system is reduced somewhat in other system than the proposed system. From this graph we can say that the accuracy of TSVM is increased which will be the best one.

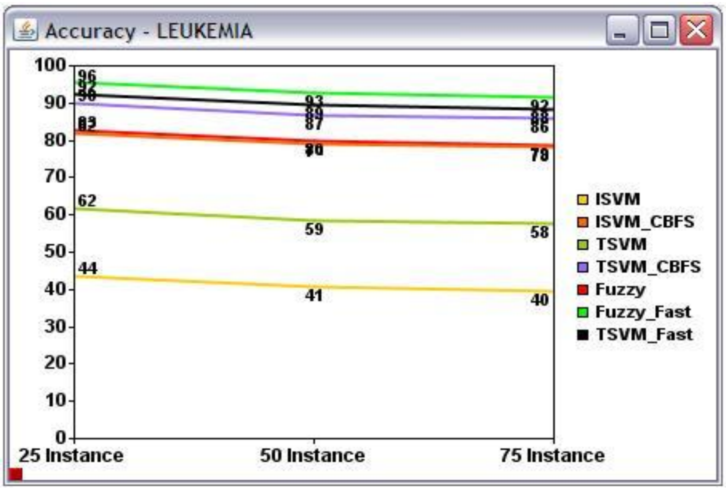

Fig.4 Accuracy Leukemia Chart

Table II and Fig. 5 represents the overall performance analysis values of all cancer datasets. Based on the results we conclude that the instances and evaluation measures are having indirect proportional relationship. That is when the instances are increased; the evaluation measures' values are decreased.

Table II. Performance Values

PERFORMANCE EVALUVATION	CANCER DATASETS			
	LEUKEMIA	MLL	SRBCT	DLBCL
ACCURACY	87.17	87.17	82.05	92.1
SENSITIVITY	90.25	90.22	84.2	94.3
SPECIFICITY	84.1	84.13	79.89	89.9

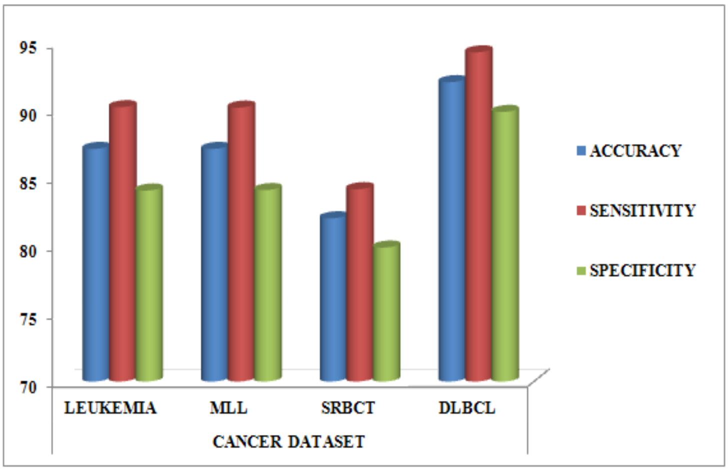

Fig. 5 Various Performance of Dataset

V. CONCLUSIONS & FUTURE WORK

In this paper mainly we focus a two technique for feature selection and classification. They are fast clustering based feature selection technique and fuzzy based transductive support vector machine (FTSVM). Compared with other different types of feature subset selection algorithms, these algorithms not only reduce the number of features, but also improves the performances of the well-known different types of classifiers. In contrast to standard approaches which make underlying assumptions about the distribution of training data, FTSVM method generates membership values based on their positions relative to the TSVM decision function. The FTSVM system has very effective rate in accuracy, specificity and sensitivity rate compared to the other systems.

The FTSVM method has only used for the selection of genes from microarray datasets. In future, this method will be extended to other density approximation tasks, and furthermore, its merits and limitations will be evaluated. Also we plan to explore different types of correlation measures and study some formal properties of feature space. Multiple features filtering stages will be used to enhance the accuracy rate of the gene classification. In addition it will reduce the time complexity rate further.

REFERENCES

[1] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D,

Brown PO, Staudt LM, *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling*, Nature, 403(6769):503-511, 2000.

[2] Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD, Sallan SE, Lander ES, Golub TR, Korsmeyer SJ, *MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia*, Nature Genetics, 30:41-47, 2002.

[3] Bandyopadhyay. S, A. Mukhopadhyay and U. Maulik, *An improved algorithm for clustering gene expression data*, Bioinformatics, 23(21):2859-2865, 2007.

[4] Bandyopadhyay. S, R. Mitra, U. Maulik, Michael Q Zhang, *Development of the human cancer microRNA network*, BMC Silence, 1(6):1-14, 2010.

[5] Bandyopadhyay. S, U. Maulik and D. Roy, *Gene identification: Classical and computational intelligence approaches*, IEEE Trans. Syst., Man, Cybern. C, 38(1):55–68, 2008.

[6] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES, *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*, Science, 286(5439):531-537, 1999.

[7] Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Ramaswamy S, Richards WG, Sugarbaker DJ, Bueno R, *Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma*, Cancer Research, 62(17):4963-4967, 2002.

[8] Hong Chai and Carlotta Domeniconi, *An Evaluation of Gene Selection Methods for Multi-class Microarray Data Classification*, Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics, 1:7-14, 2004.

[9] Hu. D Q. H. R.Yu, and Z. X. Xie, *Information-preserving hybrid data reduction based on fuzzy-rough techniques*, Pattern Recognit. Lett., 27(1):414–423, 2006.

[10] Jensen. R and Q. Shen, *Semantics-preserving dimensionality reduction: Rough and fuzzy-rough-based approaches*, IEEE Trans. Knowl. Data Eng., 16(12):1457–1471, 2004.

[11] Joachims. T, *Transductive inference for text classification using support vector machines*, Proceedings of the Sixteenth International Conference on Machine Learning (ICML '99), 1:200-209, 1999.

[12] Li T, Zhang C, Ogihara M, *A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression*, Bioinformatics, 20(15):2429–2437, 2004.

[13] Myles Hollander, Douglas A. Wolfe and Eric Chicken, *Nonparametric Statistical Methods*, NJ: Wiley, 1999.

[14] Qinbao Song, Jingjie Ni and Guangtao Wang, *A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data*, IEEE Transactions on Knowledge and Data Engineering, 25(1):1-14, 2013.

BIOGRAPHY

Dr. S. Krishnaveni completed MCA., M.Phil., Ph.D., in Computer Science and currently working as an Assistant Professor, Dept. of Computer Applications in Pioneer College of Arts and Science. Three years of experience in teaching and published thirteen papers in International Journals and also presented seven papers in various National and International conferences. Research areas includes Data mining and warehousing, Grid computing, Bioinformatics and Computer Network.

A. Sathish completed M.Sc., MCA, M.Phil., in Computer Science and currently working as an Assistant Professor, Dept. Computer Science in Maharaja College of Arts and Science. Area of research is Data mining.