Scientific Journal Impact Factor (SJIF) : 1.711
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745
ISSN (PRINT) : 2393 - 8161

Prediction of Cancer Subtypes from Microarray Data through Kernelized Fuzzy Rough Set and Association Rule Based Classification

Pheba Thomas[1], Thania Kumar[2]
PG Scholar, Computer Science and Engineering, Adi Shankara Institute of Engineering and Technology, Kalady, India
Assistant Prof, Computer Science and Engineering, Adi Shankara Institute of Engineering and Technology Kalady, India

***Abstract*—** Microarrays have now gone from obscurity to being almost ubiquitous in biological research. At the same time, the statistical methodology for microarray analysis has progressed from simple visual assessments of results to novel algorithms for analyzing changes in expression profiles. Microarray cancer data, organized as samples versus genes fashion, are being exploited for the classification of tissue samples into benign and malignant or their subtypes. In this paper, we attempt a prediction scheme that combines kernalised fuzzy rough set (KFRS) method for feature (gene) selection with association rule based classification. Biomarkers are discovered employing three feature selection methods, including KFRS. The effectiveness of the proposed KFRS and association rule based classification combination on the microarray data sets is demonstrated, and the cancer biomarkers identified from miRNA data are reported. To show the effectiveness of the proposed approach, we compare the performance of this technique with the Fuzzy Rough Set Attribute Reduction on Information Gain Ratio(FRS_GR), signal-to-noise ratio (SNR) and consistency based feature selection (CBFS) methods. Using four benchmark gene microarray datasets, we demonstrate experimentally that our proposed scheme can achieve significant empirical success and is biologically relevant for cancer diagnosis and drug discovery.

***Keywords*-** KFRS**,** Information gain ratio, microarray data, TSVM

I. INTRODUCTION

Developing simple data mining tests that allow early cancer detection is one of the top priorities in cancer research field. Cancer classification of different tumor types is of great importance in cancer diagnosis and drug discovery. Such tests will impact patient care and outcome through disease screening and early detection. Large number of gene expression/miRNA data and their diverse expression patterns indicate that they are likely to be involved in a broad spectrum of human diseases [1].The advent of microarray technology has made it possible to study the expression profiles of a large number of genes across different experimental conditions. Microarray- based gene expression profiling has shown great potential in the prediction of different cancer subtypes[4], [5], [6], [7], [8]–[12].

Major research on extending support vector machines (SVMs) to handle semi labeled data is based on the following idea: 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. In other words, this approach implements the cluster assumption for semisupervised learning, that samples

in a data cluster have identical labels. The idea was first introduced under the name of transductive SVM, but since it learns an inductive rule defined over the entire input space, the approach is referred to as semisupervised SVM (S3VM). Each cluster of samples is assumed to belong to one data class. Thus, a decision boundary is defined between clusters. A variety of semisupervised techniques have been proposed and many successful algorithms directly or indirectly assume high density within class and low density between classes, and can fail when the classes are strongly overlapping. This can be illustrated by comparing the well-known SVMs to their semisupervised extension, transductive SVM, progressive TSVM algorithm (PTSVM), transductive SVMs (TSVMs),and semisupervised SVMs (S3VMs). TSVMs and S3VMs are iterative algorithms that use SVMs to gradually search a reliable hyperplane exploiting both labeled and unlabeled samples in the training phase [13].

In this paper classification is performed using association rule. Rules are created by using Apriori algorithm which generate several if then rules[3]. In association rule mining two major rule interestingness measurements are support and confidence. Transductive support vector machine is also mentioned in this paper and accuracy is compared [13].

Selection of informative genes [22] is an important part for the analysis of microarray data. Successful feature selection has several advantages in such situations where thousands of features are involved. First, dimension reduction is employed to reduce the computational cost. Second, reduction of noises is performed to improve classification accuracy. Finally,extraction of more interpretable features or characteristics that can be helpful to identify and monitor the target diseases.In this work, we have investigated several feature selection methods namely kernelized fuzzy rough set (KFRS) [19], [20],Fuzzy Rough Set Attribute Reduction on Information Gain Ratio(FRS_GR),signal-to-noise ratio (SNR) and consistency based feature selection (CBFS) [21]. Subsequently, different tumour types are predicted based on these selected microarray biomarkers using our recently proposed transductive (semisupervised) SVM (TSVM) [14] and compared with the performances of the traditional supervised methods including SVM [23].Experimental results of the proposed method have proved to be effective based on the comparative study conducted on these microarray datasets.

II. STATE OF ART

2.1 CLASSIFICATION METHODS

2.1.1. INDUCTIVE SVM [11]

Inductive SVM (ISVM) is a general class of learning architecture originated in modern statistical learning theory.Given a training dataset, the SVM training algorithm obtains the optimal separating hyperplane in terms of generalization error. In a binary classification problem, let S = [(xi; yi)],i=1,2,.....,l be the set of training examples, where yi $\in \{\pm 1\}$is the label associated with input pattern xi. In a learning problem, the task is to estimate a function f from a given class of functions that correctly classifies unseen examples (x,y) by computing the sign(f (x)). In the case of pattern recognition, this means that given some new patterns x$\in \chi$, the classier predicts the corresponding y$\in \{\pm 1\}$.

Following nonlinear transformation, the parameters of the decision function f (x) are determined by the following minimization problem:

$\min [\Psi(w,\xi)] = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} \xi i$ (1)

subject to

$yi(\phi(xi).w + b \geq 1-\xi i , \xi i \geq 0, \quad i=1,.....,l$ (2)

where C is a user-specified, positive, regularization parameter in Eqn. (1), The variable ξi are the so called slack variables. The cost function in Eqn. (1) constitutes the structural risk, which balances empirical risk. The regularization parameter C controls this trade off.

2.1.2. TRANSDUCTIVE SVM [14]

The TSVM classifier proposed ,on the selected gene and miRNA subsets obtained by the different feature selection methods. Training

the TSVM algorithm can be roughly outlined as the following steps:

Step 1: Specify C and C^* and execute an initial

learning using the original training set to obtain a trained SVM classifier.

Step 2: Compute the decision function values of all the unlabeled samples using the trained SVM classifier. Obtain label vector of the unlabeled set. Select all the positive and negative semilabeled (transductive) samples within the margin band and add them to the original training set to obtain a hybrid training set.

Step 3: Retrain the SVM classier using this hybrid training set.Obtain the label vector of the unlabeled set. Select all the positive and negative semi labeled samples within the margin band.

Step 4: Select the common transductive samples between the previous and current transductive samples.

Step 5: Remove the previous transductive samples from the hybrid training set and add the resultant transductive set obtained from step 4.

Step 6: Repeat steps 3-5. The algorithm finishes after a finite number of iterations.

The algorithm is capable of reducing the misclassification rate of the transductive samples at each iteration through a process of successive filtering between the transductive sets which results in increased accuracy. The SVMs play the role to separate positive and negative samples, while the transductive inference successively searches more reliable discriminant function employing additional unlabeled samples. Intuitively, unlabeled patterns guide the linear boundary away from the dense regions

2.2. ATTRIBUTE SELECTION

Gene Expression data contain thousands of attribute, so attribute selection has great part in the classification purpose. So a fuzzy rough set based information gain ratio is used.

Fuzzy rough set theory has number of attribute selection approaches. A discernibility matrix method proposed by Skowron [15], in which any two objects decide one feature subset that can differentiate them. Many heuristic attribute reduction methods have been developed to support efficient attribute reduction in rough set theory. Hu and Cercone [16] proposed a heuristic attribute reduction method in which the positive region of target decision is unchanged. Grzymala-Busse [17] proposed the idea of positive region attribute reduction. Slezak [18] introduced information entropy to search reducts in rough set model.

2.2.1. FUZZY ROUGH SET BASED ATTRIBUTE REDUCTION ON INFORMATION GAIN RATIO

Fuzzy rough set theory is used to deal with real valued attribute. In real value attribute fuzzy equivalence relation is calculated instead of relation based on crisp equivalence. In crisp rough set crisp Equivalence is central ,in the case of fuzzy rough set fuzzy equivalence relation is central. If S is a fuzzy equivalence relation then S satisfies [2]

1.Reflectivity: E(a, a) = 1, ∀a ∈ A;

2. Symmetry: E(a, b) = E(b, a), ∀a, b ∈ B;

3. Transitivity: $E(a, c) \geq \min_b \{E(a, b), E(b, c)\}$. Set

Let A be a finite set and Fuzzy equivalence relation be X, denoted by a relation matrix M(E);

$$M(E)=\begin{pmatrix} e_{11} & e_{12} & \dots\dots & e_{1n} \\ e_{21} & e_{22} & \dots\dots & e_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ e_{n1} & e_{n2} & \dots\dots & e_{n3} \end{pmatrix}$$

A fuzzy decision system FDS = (U, C ∪ D, V, f),where C is the condition attribute set and D is the decision attribute.A⊆ C, ∀a ∈ C −A, the mutual information gain ratio of attribute a,

Gain Ratio(a, A, D) can be defined as

Gain Ratio(a, A, D) =Gain(a, A, D)/H({a})

= I(A ∪ {a}; D) I(A; D)/H({a})

If A =∅, Gain Ratio(a, A, D) = I({a};D)/H({a})

Then, the attribute selection based on the gain ratio is proposed in [1] by J.Dai et.al is given below

2.2.1.1.Algorithm GAIN RATIO AS FRS.(FRS_GR)

Step 1. Let A =∅;

Step 2. For all attribute a ∈ C − A, compute the significance of condition attribute a, Gain Ratio(a, A, D);

Step 3. Select the attribute which is having maximum Gain Ratio(a, A, D), store it as a; and A ← A∪ {a};

Step 4. If Gain Ratio(a, A, D) > 0, then A ← A ∪ {a}, goto Step 2,

else goto Step 5;

Step 5. The set A is the selected attributes.

2.2.2. CONSISTENCY BASED FEATURE SELECTION [21]

Consistency measure is exploited as a selection criterion that does not attempt to maximize the class separability but aims to retain the discriminatory power of the original features. A typical feature selection method has three basic steps:

1) a generation procedure to generate the next candidate subset of features;

2) an evaluation function to evaluate the candidate subset; and

3) a stopping criterion to decide when to stop .

Consistency measure is defined by inconsistency rate which is computed as follows:
Definition 1: A pattern is considered inconsistent if there exist at least two instances such that they match all but are with different class label.
Definition 2: The inconsistency count ξi for a pattern pi of feature subset is the number of times it appears in the data minus the largest number among different class labels.
Definition 3: The inconsistency rate of a feature subset is the sum,∑ξi , of all the incosistency counts over all patterns of the feature subset that appears in data divided by |U|. Correspondingly, consistency is computed as δ = (|U| −∑ξi / |U|.

2.2.3. SIGNAL TO NOISE RATIO [13]

Signal to noise ratio is calculated by using the equation

$$SNR= (m1-m2)/(\sigma1+\sigma2)$$

m1 and m2 are mean and σ1 and σ2 are standard deviations. For example take the gene expression data then calculate the SNR value of based on gene expression level. Then arrange in descending order and select top ten features. This method will enhance the accuracy of classification.

III. METHODOLOGY

The proposed method uses kernelized fuzzy rough set (KFRS) to find a set of biomarkers from the microarray datasets. Subsequently, the biomarkers are then used to distinguish to classes of samples using Fuzzy Classifier. To study the performance of the proposed method, we have used two well-known feature selection methods: fuzzy rough set on information gain ratio (FRS_GR) and consistency based feature selection (CBFS).Finally, computational and biological validations have been performed.

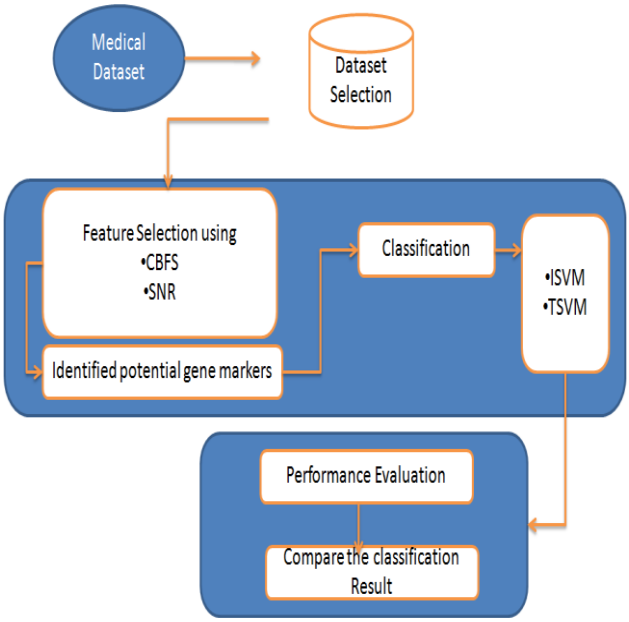

Fig 1:System Design

3.1. KERNELIZED FUZZY ROUGH SET FOR FEATURE SELECTION [24]

High level of similarity between kernel methods and rough sets can be obtained using kernel matrix as a relation [19].Kernel matrices could serve as fuzzy relation matrices in fuzzy rough sets. Taking this into account, a bridge between rough sets and kernel methods with the relational matrices was formed [19]. Kernel functions are used to derive fuzzy relations for rough sets based data analysis. In this study, Gaussian kernel approximation has been used to construct a fuzzy rough set model, where sample spaces are granulated into fuzzy information granules in terms of fuzzy T -equivalence relations computed with Gaussian kernel. The details on kernelized fuzzy rough set model is available in [19].

Formally, the forward greedy search algorithm based on Gaussian kernel approximation [20] can be written as:

Input: Sample set U = $\{z_1, z_2, ..., z_m\}$, feature set A,decision F and stopping threshold ϵ

Output: reduct red

Step 1: Initialize red to an empty set and β to 0.

Step 2: For each attribute ai $\in$ A – red ,compute

$\beta_i = \beta_{\{ai\}} \cup red$

Step 3: Find the maximal β_i and the corresponding attribute a_i

Step 4: Add attribute a_i to red if it satisfies

$\beta_i - \beta_{red}(F) > \varepsilon$

Step 5: Assign β_i to β_{red}

Step 6: Repeat steps 2-5 while red $\neq$ A

Step 7: Return red

Initially, the algorithm starts with an empty set of attribute.Subsequently, it evaluates the remaining attributes at each iteration and selects feature producing the maximal fuzzy dependency. Algorithm for the computation of dependency with Gaussian kernel is available in [20]. The algorithm terminates when adding any of the remaining attributes does not satisfy step 4 in the above algorithm. The output of the algorithm is a reduced feature set.

The fuzzy dependency (F) can be computed as follows:

Input: Sample set U = {z1. z2,...., zm}, feature set A, decision F and parameter δ

Output: dependency β of F to A

Step 1: $\beta_A(F) \leftarrow 0$

Step 2: i = 1 to m

Step 3: find the nearest sample xi of zi with a different class

Step 4: $\beta A(F) \leftarrow \beta A(F) + \sqrt{1 - \left[\exp\left(-\left\{||zi - xi||^2\right\}/\delta\right)\right]^2}$

Step 5: return $\beta_A(F)$

The algorithm will remove those features from the data which would receive low dependency values.

3.2 CLASSIFICATION BASED ON ASSOCIATION RULE

One of the important component of data mining is association rule .Main objective of association rule is to discover all the co-occurrence relations called association. Association rule can contain more than one item in predecessor and resultant of the rule. There are two constraints for every rule support and confidence .Support is measure of statistical significance and confidence is measure of goodness. Association rule is in the form ,x→ y where x or y is items in itemset.

$$support = \frac{(X \cup Y).count}{n}$$

$$confidence = \frac{(X \cup Y).count}{X.count}$$

Association rule was created using Apriori algorithm .

3.2.1.Apriori algorithm

Apriori algorithm is one of best known algorithm in datamining.It is used to create large number of item set .It was proposed by Rakesh agarwal et.al.[3] This algorithm works in two steps.

1)First of all it will generate all frequent itemset. Frequent itemset will be having support more than minimum support.

2)From the frequent itemset confident association rule is generated, which is having more confident than minimum confident.

Apriori algorithm depends on downward closure property to create all frequent itemset.

3.2.1.1.Algorithm Apriori

In Apriori algorithm. The first pass will counts item occurrences to find the large 1-itemsets. First,In (k-1)th pass the large itemset L_{k-1} is found used to generate the candidate itemsets Ck, using the apriori_can_gen function described in Section 2 the

Let L_k Set of large k-itemsets (those with minimum support).There will be two fields for for each member they are i)itemset and ii) support count.

Let Ck bethe Set of candidate k-itemsets

Two fields in each member are i) itemset and ii) support count t be the transaction

The algorithm proposed by rakesh agarwal is given below[2]

Algorithm

1)Assign L_1 = {large 1- itemsets}
2) for (p= 2; $L_{k-1} \neq 0$; k++) do begin
3) S_k = apriori-gen(L_{k-1}); // candidates which are new
4) for every transactions t ∈ D do begin
5) Ct = subset(Ck , t); // Candidates contained
6) for every candidates c ∈ Ct do
7) c.count++;
8) end
9) Lk = {c ∈ C_k | c.count ≥ minsup }
10) end
11) Answer = $\bigcup_k L_k$

Algorithm Apriori

3.2.1.2 Apriori Candidate Generation

For the set of all large (k - 1)-itemsets the apriori_can_gen function takes L_{k-1} . It provide a superset for the set of all large k-itemsets.

The function apriori_can_gen works as follows.

- In the join step , join L_{k-1} with L_{k-1}:
- insert into C_k
- Then select m.item$_1$, m.item$_2$, ..., m.item$_{k-1}$, n.item $_{k-1}$ from L_{k-1} m, L_{k-1} n

Where $m.item_1 = n.item_1, \ldots .., m.item_{k-2} = n.item_{k-2}, m.item_{k-1} < n.item_{k-1}$;

- Delete all item sets $c \in C_k$ [2]

So by using this algorithm rules are generated and these generated rules are used for training and testing and calculate the classification accuracy.

REFERENCES

[1] E. Berezikov, E. Cuppen, and R.H.A. Plasterk, ``Approaches to microRNA discovery," Nature Genet., vol. 38, pp. S2S7, May 2006.

[2] Jianhua Dai, Qing Xu ," Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification" Applied Soft Computing 13 (2013) 211–221

[3] Rakesh Agrawal and Ramakrishnan Srikant Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.

[4] S. Bandyopadhyay, A. Mukhopadhyay, and U. Maulik, "An improved algorithm for clustering gene expression data," Bioinformatics, vol. 23,no. 21, pp. 2859–2865, 2007.

[5] S. Bandyopadhyay, R. Mitra, and U. Maulik, "Development of the human cancer microRNA network," BMC Silence, vol. 1, no. 6, 2010.

[6] A.J.Gentles,S.K.Plevritis, R. Majeti, and A. A. Alizadeh, "Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloid leukemia," JAMA—J. Amer. Med. Assoc.,vol. 304, no. 24, pp. 2706–2715, 2010.

[7] H. K. Kim, I. J. Choi, C. G. Kim, A. Oshima, and J. E. Green, "Gene expression signatures to predict the response of gastric cancer to cisplatin and fluorouracil," J. Clin. Oncol., vol. 27, no. 15s, 2009.

[8] U. Maulik, A. Mukhopadhyay, and S. Bandyopadhyay, "Combining pareto-optimal clusters using supervised learning for identifying co-expressed genes," BMC Bioinformat., vol. 10, no. 27, 2009.

[9] U.Maulik, "Analysis of gene microarray data in soft computing framework," Appl. Soft Comput., vol. 11, no. 6, pp. 4152–4160, 2011.

[10] A. Mukhopadhyay, S. Bandyopadhyay, and U. Maulik, "Multi-class clustering of cancer subtypes through SVM based ensemble of pareto-optimal solutions for gene marker identification," PLoS ONE, vol. 5,no. 11, pp. 1–14, 2010.

[11] U.Maulik,S.Bandyopadhyay,andA.Mukhopadhyay,MultiobjectiveGenetic Algorithms for Clustering: Applications Data Mining and Bioinformatics. New York: Springer-Verlag, 2011.

[12] U. Maulik and A. Mukhopadhyay, "Simulated annealing based automatic fuzzy clustering combined with ANN classification for analyzingmicroarray data," Comput. Oper. Res., vol. 37, no. 8, pp. 1369–1380, 2010.

[13] U. Maulik, A. Mukhopadhyay, and D. Chakraborty, "Gene-expression based cancer subtypes prediction through feature selection and transductive SVM," IEEE Trans. Biomed. Eng., vol. 60, no. 4, pp.1111–1117, 2013.

[14] U. Maulik and D. Chakraborty, "Fuzzy Preference Based Feature Selection and Semisupervised SVM for Cancer Classification," IEEE Transactions on nanobioscience, vol. 13, no. 2, June 2014.

[15] A. Skowron, Extracting laws from decision tables: a rough set approach, Computational Intelligence 11 (1995) 371–388.

[16] X.H. Hu, N. Cercone, Learning in relational databases: a rough set approach, Computational Intelligence 11 (1995) 323–338.

[17] J. Grzymala-Busse, An algorithm for computing a single covering, in: Managing Uncertainty in Expert Systems, Kluwer Academic Publishers,1991, p. 66.

[18] D.Slezak, Foundations of entropy-based Bayesian networks: theoretical results & rough set based extraction from data, in: Proceedings of the 8th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, 2000, pp. 248–255.

[19] Q. Hu, D. Yu, W. Pedrycz, and D. Chen, ``Kernelized fuzzy rough sets and their applications," IEEE Trans. Knowl. Data Eng., vol. 23, no. 11,pp. 16491667, Nov. 2011.

[20] Q. Hu, L. Zhang, D. Chen, W. Pedrycz, and D. Yu. Gaussian Kernel Based Fuzzy Rough Sets: Model, Uncertainty Measures and Applications.[Online].Available:http://www4.comp.polyu.edu.hk/

[21] M.Dash andH. Liu, ``Consistency-based search in feature selection," Artif.Intell., vol. 151, nos. 12, pp. 155176, Dec. 2003.

[22] J. C. Rajapakse and P. A.Mundra, ``Multiclass gene selection using Pareto-fronts," IEEE Trans. Comput. Biol. Bioinformat., vol. 10, no. 1, pp. 8797,Jan./Feb. 2013.

[23] V. N. Vapnik, Statistical Learning Theory. New York, NY, USA: Wiley,1998.

[24] U. Maulik and D. Chakraborty, "Identifying Cancer Biomarkers From Microarray Data Using Feature selection and Semisupervised learning," IEEE Journal of Transactions in Health and Medicine, vol. 13, no. 2, Dec 2014.