INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE AND ENGINEERING TECHNOLOGY (IJRASET)

Cancer prediction Based on Gene Expression data Through Association Rule Based classification and Fuzzy Rough Set Attribute Reduction on Information Gain Ratio

Neethu Innocent[#1], Mathew Kurian[*2]

[#]*PG Studen,t Computer Science Department, Karunya University*

**Assistant Professor,Computer Science Department, Karunya University*

Abstract— ***Data mining hast vast number of application in the area of medical science. This paper mainly aim to predict cancer type based on gene expression data. For attribute selection Information gain ratio on fuzzy rough set theory is used. From the selected gene classification is performed by using association rule ,rule is created by using apriori algorithm. These algorithm is compared with some algorithm and accuracy is evaluated.***

Keywords— ***Fuzzy Rough set, Information gain ratio, Apriori algorithm, Transductive support vector machine***

I. Introduction

Data mining has tremendous application in the area of medical field .Nowhere days cancer is hunting many people life , so accurate prediction of cancer subtype will help for correct diagnosis of disease. Gene expression provides the activation level of each gene of an organism at particular period of time. Micro array technology is used to study different expression profile of large number of gene . In this paper gene expression data of cancer is used for classification of cancer types .[2][3][4]

There are several methods for classification in data mining .By using support vector and semi supervised classification methods like Transductive support vector machine, Semi supervised SVM, Low density separation approach etc .

In this paper classification is performed using association rule. Rules are created by using Apriori algorithm which generate several if then rules[6]. In association rule minig two major rule interestingness measurements are support and confidence. Transductive support vector machine is also mentioned in this paper and accuracy is compared [2].

Gene expression dataset is a huge data. So before classification the dimensionality of data must reduced to improve performance and accuracy. For selection criteria information gain ratio based on fuzzy rough set attribute reduction is used Consistency based feature selection is also used for comparison purpose. The flowchart of the steps which perform in this paper is given in fig;1.

In this paper section II mentioned about Attribute selection it has two methods fuzzy rough set based information gain ratio and consistency based feature selection. Section III explained about two classification method which are association rule and by using Transductive procedure. Experiment and result is proved in section IV and concluded in section V.

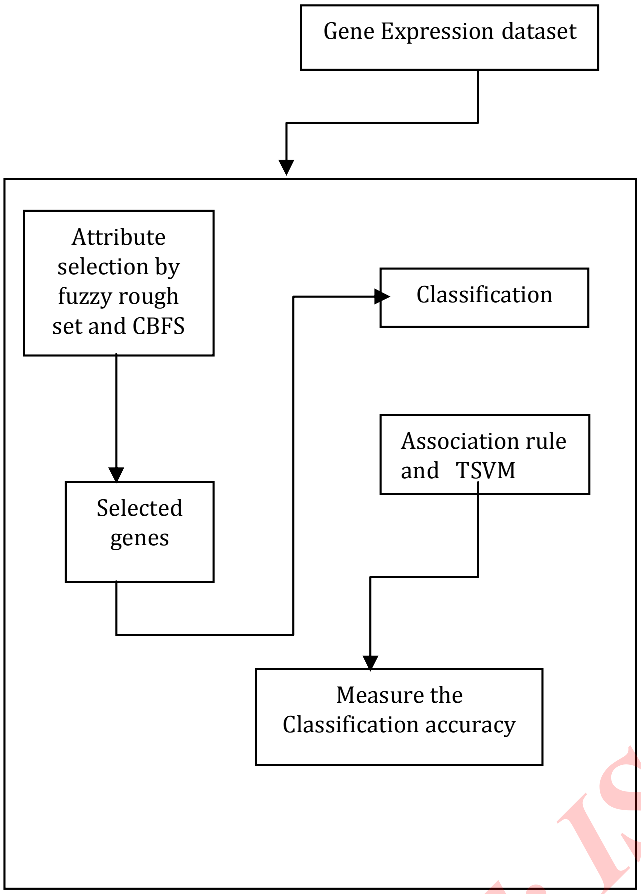
Gene Expression dataset
Attribute selection by fuzzy rough set and CBFS
Classification
Association rule and TSVM
Selected genes
Measure the Classification accuracy

Fig1:Flow chart

II.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 [9] , 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 [10] proposed a heuristic attribute reduction method in which the positive region of target decision is unchanged. Grzymala-Busse [11] proposed the idea of positive region attribute reduction. Slezak [12] introduced information entropy to search reducts in rough set model. In this paper it focus on information gain based attribute selection approach.

A.Fuzzy Rrough 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 [1]

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);

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

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

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.
```

B.Consistency based Feature selection

To increase the accuracy of classification system relevant attribute of huge data is selected for this consistency based feature selection also used .Consistency measure is calculated by using inconsistency rate.[1][8][5]

1)If two instances matches all but they are from different classes then that pattern is considered to be inconsistent

2) The inconsistency count is calculated for feature subset is the number of times it appears in the data minus the largest number among different class labels.

3)The feature subset inconsistency rate is calculated .It is the sum of all inconsistency count of overall pattern and is divided by total number of instances.

III.CLASSIFICATION

A.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. Association rule was created using Apriori algorithm.

$$support = \frac{(X \cup Y).count}{n} \qquad (1)$$

$$confidence = \frac{(X \cup Y).count}{X.count} \qquad (2)$$

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.[6] 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.

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[6]

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 = $\bigcup_k L_k$
```

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.

Join L_{k-1} with L_{k-1}:

insert into C_kThen 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$_1$, ..., m.item$_{k-2}$ = n.item$_{k-2}$, m.item$_{k-1}$ < n.item$_{k-1}$;

Delete all item sets c ∊ C_k [6]

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

B.Transductive Support vector Machine

TSVM use both labeled and unlabeled samples in training phase which is an iterative algorithm. Ujwal malik et.al proposed a transductive procedure in which transductive sample is selected through a filtering process of unlabeled data and an algorithm is proposed[2] [5].

Input of this algorithm will both labeled and unlabeled samples. Algorithm starts with training the SVM classifier which is having a working set T(0).Working set will be equivalent to labeled set . The unlabeled data which drop into the margin will have more details, in which some data drop into negative side and is called negative transductive sample and unlabeled data drop into positive side is called positive transductive samples.Informative samples with accurate labeling and samples which are near to margin will be selected and some which is residing in upper side and in lower side will be assigned as +1and -1 respectively. Selected transductive sample is added to the training set. So by using this method unlabeled samples also added to the training .

IV.EXPERIMENTS AND RESULT

Gene Expression dataset used are

1. Lung cancer:This data set contain 12600 genes and 203 samples It has four subtype [14] adenocarcinoma (AD): 139 samples , normal lung (NL): 17samples, small cell lung cancer (SMCL): 6 samples, squamous cell carcinoma (SQ): 21 samples , pulmonary carcinoid (COID): 20 samples .

2.postrate cacer: This dataset contain 102 samples and two subtypes . normal tissue (normal): 50 samples, prostate tumor (tumor): 52 samples

TABLE I

ACCURACY OF DIFFERENT ALGORITHM

Cancer types	**TSVM + CBFS**	**TSVM+ FRS_GR**	**AR+ CBFS**	**AR+ FRS_GR**
Lung cancer	90%	80%	83.5%	92.5%
Postrate cancer	92%	82.5%	87.5%	94.3%

V. Conclusions

In this paper cancer types is predicted by using two method for attribute selection fuzzy rough set based information gain ratio is used and it is compared with consistency based feature selection. For classification association rule is used apriori algorithm is used to create rule and is compared with Transductive support vector machine .

After comparison combination of fuzzy rough set based information gain ratio and Association rule has more accuracy than other algorithm.

References

[1] 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

[2] Ujjwal MauliK, Anirban Mukhopadhyay and Debasis Chakraborty "gene-expression-based cancer subtypes prediction through feature selection and transductive SVM" IEEE transactions on biomedical engineering, vol. 60, no. 4, april 2013

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

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

[5] Neethu Innocent , Mathew Kurian ,"Survey on semi supervised classification methods and feature selection". International Journal of Research in Engineering and Technology .pISSN: 2321-7308(2014)

[6] 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.

[7] Q. Hu, D. Yu, Z. Xie, Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognition Letters 27 (2006) 414–423.

[8] M. Dash andH. Liu, "Consistency based search in feature selection," Artif.Intell., vol. 151, pp. 155–176, 2003.

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

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

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

[12] 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 Uncertaintyin Knowledge-Based Systems, 2000, pp. 248–255.

[13]Available:http://www.biolab.si/supp/bi-cancer/projections/ index.htm