Volume: 04 Issue: 01 June 2015 Page No.5-9
ISSN: 2278-2419

Correlations Based Rough Set Method for Diagnosis and Drug Design using Microarray Dataset

Sujata Dash[1], Bichitrananda Patra[2]
[1]Department of Computer Science, North Orissa University,Baripada, Odisha, India.
[2]Department of Computer Science & Engineering, KMBB College of Engineering & Technology,
Bhubaneswar, Odisha, India.
Email:sujata_dash@yahoo.com, bnpatra@gmail.com

Abstract-Databases containing huge amount of genetic information about diseases related to cancer are beyond our capability to analyze and predict the discriminative characteristics of the genes involved. But, this kind of analysis helps to find the cause and subsequent treatment for any disease. In this work, a hybrid model has been developed combining the characteristics of Rough set theory (RST) and Correlation based feature subset (CFS) selection technique which is capable of identifying discriminative genes from the microarray dataset. The model is tested with two publicly available multi-category microarray dataset such as Lung and Leukemia cancer. The study reveals that Rough set theory (RST) is capable of extracting predictive genes in the form of reducts from the subset of genes which are highly correlated with the class but having low interaction with each other. The performance of the model has been evaluated via three learning algorithms using 10-fold cross validation. This experiment has established that the hybrid supervised correlation based reduct set (CFS-RST) method is able to identify the hidden relationships among the genes which cause diseases as well as help to automate medical diagnosis. Finally, the functions of identified genes are analysed and validated with gene ontology website DAVID which shows the relationship of genes with the disease.

Keywords - gene selection, correlation, rough set theory, reducts microarray dataset

I. INTRODUCTION

With the development of microarray technology, DNA microarrays with millions of genes have been obtained. Finding the genes which are related to cancer is significant to medical treatment. There are various kinds of cancers. Each type of cancer may connect to different genes. Distinguishing classes of cancer based on gene expression levels has great importance on cancer diagnosis [1]. There are a large number of genes in the gene expression data sets, but only a few of them are essential to the classification of a certain cancer. How to extract the relevant genes to a certain cancer becomes a key issue for cancer diagnosis. In fact the hidden relationships exist within the dataset can provide biological information which can be used as a diagnostic tool to identify the marker genes. Eventually, these marker genes help in finding accurate and appropriate treatment procedure for a specific disease. This vital biological information can be obtained employing suitable soft computing tools such as rough set theory, fuzzy theory, genetic algorithm [2], neural network etc to analyze the hidden relationship exist in the dataset. Usually, microarray datasets contain certain amount of superfluous or inconsistent data which represent the genetic profile of normal and diseased tissues of patients. Efficient and effective gene extraction methods need to be devised to handle large amount of data by most techniques. Growing interest in developing methodologies that are capable of dealing with imprecision and uncertainty is apparent from the large scale research that are currently being done in the areas related to fuzzy [3] and rough sets [4]. The success of rough set theory is due to three aspects of the theory. First, only hidden facts in the data are analyzed. Second, additional information about the data is not required for data analysis. Third, it finds a minimal knowledge representation for data.

In practice, filtering and classification algorithms are widely adopted to analyze gene expression data[5], [6].In this paper, we focus on marker gene identification using gene expression data, which is a hot topic in recent years and has received general attention by many biological and medical researchersFeature/gene selection algorithms typically fall into two categories: feature ranking and subset selection. Feature ranking ranks the features by a metric and eliminates all features that do not achieve an adequate score. Subset selection algorithms can be broken into filters, wrappers and embedded. (1)Filter type methods are essentially data pre-processing or data filtering methods. Features are selected based on the intrinsic characteristics, which determine their relevance or discriminative powers with regard to the targeted classes. Simple methods based on mutual information [6], statistical tests (t-test, F-test) have been shown to be effective [1], [7].They also have the virtue of being easily and very efficiently computed. In filters, the characteristics in the feature selection are uncorrelated to that of the learning methods; therefore they have better generalization property. (2)In wrapper type methods, feature selection is "wrapped" around a learning method. The usefulness of a feature is directly judged by the estimated accuracy of the learning method. One can often obtain a set with a very small number of non-redundant features, which gives high accuracy, because the characteristics of the features match well with the characteristics of the learning method. Rough set theory (RST) [8], [3, 4, 9] has been used as an approximation tool to discover data dependencies and to reduce the number of attributes contained in an inconsistent and ambiguous dataset using the data alone, requiring no additional information [10].

Over past few decades, RST has become a topic of great interest to researchers and has been applied to many domains. Given a dataset with discretized attribute values, it is possible to find a subset (termed a reduct) of the original attributes using RST that are the most informative and all other attributes can be removed from the dataset with minimal information loss. A quick search of biological literatures show that rough sets are still seldom used in bioinformatics problems and rarer is the use in multi-class microarray dataset. To carry out this work, we have used two multi-category microarray datasets collected from the website http://www.gems-system.org. This paper is organized in the following manner. Section 2 introduces the concept of rough set theory (RST) and correlation based feature subset selection technique. Section 3 presents the proposed hybrid supervised correlation based rough set method (CFS-RST). Section 4 describes the application of induction algorithms particularly the use of decision algorithm on microarray dataset along with the experimental and validation of generated rules using DAVID. In the concluding section, a discussion on the novel approach is followed by conclusion and references.

II. Rough Set Theory

In rough set theory, a information system is denoted by $I=(U, A\cup\{d\})$, where U is the universe with a non-empty set of finite objects, A is a non-empty finite set of conditions attributes, and d is the decision attribute (such a table is also called decision table), $\forall a \in A$ there is a corresponding function $f_a: U \rightarrow V_a$, where $V_a$is the set of values of a. If P ⊆ A, there is an associated equivalence relation:

$$IND(P) = \{(x, y) \in U \times U | \forall a \epsilon P, f_a(x) = f_a(y)\} \qquad (1)$$

The partition of U generated by IND(P) is denoted U/P . If (x,y) ∈ IND(P) , then x and y are indiscernible by attributes from P . The equivalence classes of the P-indiscernability relation are denoted by $[x]_p$. Let X ⊆ U, the P-lower approximation $\underline{P}X$ and P-upper approximation $\overline{P}X$ of set X can be defined as:

$$\underline{P}X = \{x \in U \mid [x]\ p \subseteq X\} \qquad (2)$$
$$\overline{P}X = \{x \in U \mid [x]p \cap X \neq \phi\} \qquad (3)$$

Let $P, Q \subseteq A$ be equivalence relations over U, then thepositive, negative and boundary regions can be defined as :

$$POS_p\ (Q) = \cup_{x \in U/Q} \underline{P}X \qquad (4)$$
$$NEG_p(Q) = U - \cup_{X \in U/Q} \overline{P}X \qquad (5)$$
$$BND_p\ (Q) = \cup_{X \in U/Q} PX - \cup_{X \in U/Q} PX \qquad (6)$$

The positive region of the partition U/Q with respect to P, $POS_P(Q)$ is the set of all objects of U that can be certainly classified to blocks of the partition U/Q by means of P.Q depends on P in a degree k $(0\leq k\leq 1)$ denoted by $P \underset{k}{\Rightarrow} Q$

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|U|} \qquad (7)$$

Where P is a set of condition attributes, Q is the decision, and $\gamma_P(Q)$ is the quality of classification. If k=1, Q depends totally on P; if 0<k<1, Q depends partially on P; and if k=0 then Q does not depend on P. The goal of attribute reduction is to remove redundant attributes so that the reduced set provides the same quality of classification as the original. The set of all reducts is defined as:

$$Red(C) = \{R \subseteq C | \gamma_R(D) = \gamma_C(D), \forall B \subset R, \gamma_B(D) \neq \gamma_C(D)\} \qquad (8)$$

A dataset may have many attribute reducts. The set of all optimal reducts is:

$$Red(C)_{min} = \left\{R \in Red \middle| \forall R^{`} \in Red, |R| \leq \left|R^{`}\right|\right\} \qquad (9)$$

III. CORRELATION BASED ROUGH SET GENE METHOD

Our learning problem is to select high discriminating genes for cancer classification from gene expression data. We may formalize this problem as a decision system I= $(U, A\cup\{d\})$, where universe U = {x_1, x_2 , …..,x_m} is a set of tumors. The conditional attributes set A = $\{g_1 , g_2 ,……, g_n\}$ contains each gene, the decision attribute $D = \{d\}$ corresponds to class label of each sample. Each attribute $g_i \in A$is represented by a vector g_i = {$x_{1,i}$,$x_{2,i,\ldots\ldots}$,$x_{m,i}$}, i=1,2,……,n, where x $_{k,i}$is the expression level of gene i at sample k, $k=1,2,\ldots\ldots m$. In thousands of genes many are highly correlated, this “redundancy” will increase the computational cost and at the same time decrease the accuracy of classification. Thus correlation based feature selection (CFS) is applied to decrease the dimensions of gene space as the first step.CFS evaluates a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them [11].

$$CFS_S = \frac{k\overline{r}_{cf}}{\sqrt{k + k(k-1)\overline{r}_{ff}}} \qquad (\ 10\ )$$

where CFS_S is the score of a feature subset S containing k features, ¯rcf is the average feature to class correlation (f ∈S), and ¯rff is the average feature to feature correlation. The distinction between normal filter algorithms and CFS is that while normal filters provide scores for each feature independently [11], CFS presents a heuristic “merit” of a feature subset and reports the best subset [12] it finds.The above operation can be seen as a filter of the original attribute set and reduct is then constructed from the filtered attribute set by evaluating the degree of dependency which leads to the decision attribute.

A. Supervised Correlation-based-Rough set(CFS-RST) Algorithm

The supervised CFS-RST algorithm given in Algorithm 1 to calculate the reduct from the subset [13][14] generated from CFS filter. It has two parameters, conditional attribute and decision attribute and its evaluation of degree of dependency value leads to the decision attribute. It starts off with an empty set and adds in turn, one at a time, those attributes that result in the greatest increase in the rough set dependency metric, until this produces its maximum possible value for the dataset. According to the algorithm, the dependency of each attribute is calculated and the best candidate is chosen.Algorithm 1.The supervised CFS-RST algorithm.

CFS-RST (CFS_S, D)

$CFS_{S=}\{g_1,g_2,\ldots,g_k\}$, the set of all conditional features contain each gene.
$D=\{d\}$, the set of decision features corresponds to class label of each sample.

```
$R \leftarrow \{\ \}$
do
$T \leftarrow R$
$\forall\ x \in (CFS_S - R)$
If $Y_{R\sqcup\{x\}}(D) > Y_T(D)$
$T \leftarrow R \sqcup \{x\}$
$R \leftarrow T$
until $Y_R(D) = = Y_C(D)$
return R
```

IV. CLASSIFICATION

The data mining tool WEKA [15] is an open source java based machine-learning workbench that can be run on any computer in which a java run time environment is installed. It brings together many machine learning algorithm and tools under a common frame work. The tool is used to perform benchmark experiment. Three learning algorithms were employed for the classification of the data, 1-KNN, Naïve Bayes Updateable and J48.

A. Experimental Results

Two public multi-category cancer microarray datasets such as leukemia cancer and lung cancer dataset are used to evaluate the performance of the proposed method. The Leukemia dataset consists of 72 samples, 3 classes and 11226 genes, including 25 ALL type of Leukemia, 20 MLL and 27 AML type of Leukemia data. The information of the dataset contains names of dataset, number of samples, number of classes and number of attributes, which are given in Table 1.The samples are taken from 63 bone marrow samples and 9 peripheral blood samples.The Lung cancer data set consists of 203 samples, 5 classes and 12601 genes which consists of 139 AD, 17 NL, 6 SMCL, 21 SQ and 20 COID. Both the data sets have been obtained from http://www.gems-system.org . A simple method introduced in [16] has been used to discretize the domain of each attribute because rough sets methods require discretized input. After using correlation based feature selection (CFS) and a heuristic search in the dataset, 150 genes are left in Leukemia data set and 473 genes are left in lung cancer data set. In the next step, the filtered data sets undergo attribute reduction process in which only 6 and13 genes are left in the reduct of Leukemia and lung cancer datasets respectively which is shown in table 1. Three different classification algorithms: 1-KNN, J48 and Naive Bayes Updateable are employed to evaluate the discrimination power of the obtained genes, and 10-fold cross validation method, which is a widely used process for measuring the performance of classification, has been used. A summary of the experimental results is shown in table 2, 3, 4 and 5. The pruned decision tree obtained from J48 for both the datasets help us to identify the marker genes responsible for causing the disease.The pruned decision tree obtained from J48 classifier for Lung and Leukemiacancer dataset contain the following marker genes. The function of the genes for both the dataset are studied from DAVID and shown in table 6 and 7.

TABLE I DATASET INFORMATION

Dataset	# classes	# samples	# genes	CFS	CFS-RST
Leukemia	3	72	11226	150	6
Lung	5	203	12600	473	13

TABLE II 10-FOLD CROSS VALIDATION ACCURACY OF LEARNING ALGORITHMS ON REDUCED (CFS-RST) LEUKEMIA DATASET (%)

Classifier	**Classification accuracy for each class**			**Overall classification accuracy**
	ALL	***MLL***	***AML***	
1-knn	24/24	20/20	28/28	100
Naive Bayes Updateable	24/24	17/20	28/28	94.7368 (54/3)
J48	23/24	20/20	28/28	98.2456(56/1)

TABLE III 10-FOLD CROSS VALIDATION ACCURACY OF LEARNING ALGORITHMS ON UNREDUCED LEUKEMIA CANCER DATASET (%)

Classifier	**Classification accuracy for each class**			**Overall classification accuracy**
	ALL	*MLL*	*AML*	
1-knn	17/24	17/20	25/28	81.9444 (59/13)
Naive Bayes Updateable	23/24	19/20	27/28	95.8333 (69/3)
J48	23/24	16/20	21/28	83.3333 (60/12)

TABLE IV 10-FOLD CROSS VALIDATION ACCURACY OF LEARNING ALGORITHMS ON REDUCED (CFS-RST) LUNG DATASET (%)

Classifier	**Classification accuracy for each class**					**Overall classification accuracy**
	AD	***SQ***	***NL***	***SMCL***	***COID***	
1-knn	139/139	17/17	6/6	21/21	20/20	100
Naive Bayes Updateable	134/139	17/17	6/6	21/21	20/20	97.5369 (198/5)
J48	137/139	16/17	4/6	19/21	20/20	96.5517 (196/7)

TABLE V 10-FOLD CROSS VALIDATION ACCURACY OF LEARNING ALGORITHMS ON UNREDUCED LUNG CANCER DATASET (%)

Classifier	**Classification accuracy for each class**					**Overall classification accuracy**
	AD	***SQ***	***NL***	***SMCL***	***COID***	
1-knn	134/139	12/17	2/6	16/21	18/20	89.6552 (182/21)
Naive Bayes Updateable	136/139	13/17	4/6	17/21	20/20	94.5813 (192/11)
J48	134/139	16/17	4/6	16/21	19/20	93.1034 (189/14)

The classification performance of the reduced dataset has outperformed the performance of dataset in which the proposed reduction algorithm (CFS-RST) has not been applied. We obtained 100%, 94.7368% and 98.2456% overall classification accuracy on reduced leukemia data set and 100%, 975369% and 96.5517% classification accuracy on reduced lung dataset employing 1-KNN, Naïve Bayes Updateable and J48 as classification algorithms respectively, which are all compared and found superior to [17]][18][19]. The classification accuracy for all the classes and overall accuracy of unreduced dataset of Leukemia and lung cancer are shown in table 3 and 5. The pruned decision trees obtained from J48 are have also been analysed to study the functions and behaviour of the significant genes. The above results indicate that our method has successfully achieved its objectives: automatic gene selection for predicting the class of new object. The classification accuracy of lung cancer data set is higher than leukemia data set, the reason maybe because the scale of lung data set is larger.

V. DISCUSSION

The decision table in rough set is formed using the attributes and samples of the datasets. The sample is considered as cancerous or normal basing on the conditions and expressions of the genes. This decision table is used as a training dataset which is used to find the dependencies hidden among the genes. The values of the attributes are real but it is considered as integer for RST analysis. To derive the reduct set and core from a huge number of attributes, we have used best first search to find the optimal feature subset.CFS-RST method can select few dominant genes from large number of genes. These dominant genes can be considered as the responsible genes for Leukemia and Lung cancer. The list of responsible genes for both the data sets is shown in Table 6 and 7. These genes are the indispensible core genes which are common in all reducts. CFS-RST successfully reduced the decision table and extracted 6 and 13 genes from huge amount of genes. So, rough set can be used to design an automated disease diagnosis system for microarray datasets.

TABLE VI FUNCTIONAL CLASSIFICATION OF MARKER GENES FOR LEUKEMIA.

Gene identifiers	Gene Name	Gene Symbol	Gene functions obtained from DAVID
1325_at	SMAD family member 1	SMAD1	This gene encodes a protein Migrates to the nucleus when complexed with SMAD4. Highest expression seen in the heart and skeletal muscle.
31481_s_at	Thymosin beta 10	TMSB10	This gene plays an important role in the organization of the cytoskeleton. Belongs to the thymosin beta family.
41672_at	F-box and leucine-rich repeat protein 14	FBXL14	The protein encoded by this gene is a member of a Substrate-recognition component of the SCF .
31343_at	Interleukin 1 receptor antagonist	IL1RN	Features include increased urine protein and declining kidney function.,function:Inhibits the activity of IL-1 by binding to its receptor.

Validations of the obtained results. - The validation of the result is explained in two ways i.e., mathematical validation and functional classification of genes. The mathematical validation is based on the induced rules generated from the reducts obtained from CFS-RST reduct algorithm and classification of whole dataset on the basis of that generated rules which have only few responsible genes. The accuracy of prediction of the diseases is verified by applying pruned tree generated by decision tree J48 classifier on the datasets. The experiment shows that the datasets contain only few marker genes which classify the entire datasets.

TABLE VII FUNCTIONAL CLASSIFICATION OF MARKER GENES FOR LUNG.

Gene identifiers	Gene Name	Gene Symbol	Gene functions obtained from DAVID
35693_at	hippocalcin-like 1	hpcal1	This gene May be involved in the calcium-dependent regulation of rhodopsin phosphorylation., miscellaneous.
41453_at	discs, large homolog 3 (Drosophila)	DLG3	This gene plays an important role in mental retardation.
41038_at	neutrophil cytosolic factor 2	Ncf2	The protein encoded by this gene is a member of a NCF2, NCF1, and a membrane bound cytochrome.
36032_at	heat shock protein family B	Hspb11	These data represent a suitable target for modulating cell death pathways directly correlated with the histological grade of brain tumors.
37754_at	galactoside-binding	LGALS3BP	Promotes intergrin-mediated cell adhesion. May stimulate host defense against viruses and tumor cells.
1582_at	Carcinoembryonicanigen related cell adhesion molecule 5	CEACAM	Cell surface glycoprotein that plays a role in cell adhesion and in intracellular signaling. Found in adenocarcinomas of endodermally derived digestive system epithelium and fetal colon.
41288_at	calmodulin 3 (phosphorylase kinase, delta);	CALM1, CALM2, CALM3	Calmodulin mediates the control of a large number of enzymes and other proteins by Ca(2+).

VI. CONCLUSION

In this work, a new supervised CFS-RST (Correlation based rough set algorithm) algorithm using rough set theory is proposed. The method attempts to calculate a reduct using the minimal subset obtained from CFS filter. It starts off with an empty set and adds in turn, one at a time, those attributes that result in the greatest increase in the rough set dependency metric, until this produces its maximum possible value for the dataset. Two well known public multi class datasets are used to test the performance of this novel method. High prediction accuracies have been achieved through 10 fold cross validation method, this suggests that our method can select marker genes for cancer classification which is validated through DAVID gene ontology website. Rough set approach holds a high potential to become a useful tool in bioinformatics. In future, the proposed method can be compared with a hybrid PLS-quick reduct algorithm using multi-class microarray dataset for cancer classification. The biological relevance of the method is studied by finding the actual functional classification of the marker genes captured from the classification results.

REFERENCES

[1] Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science, Vol. 286, No.5439,pp 531–537, 1999.

[2] Dash, S.: Hill-climber Based Fuzzy-Rough Feature Extraction with an Application to Cancer Classification, Journal of Network and Innovative Computing (JNIC), ISSN 2160-2174 (HIS 2013 Special Issue) (In Press)

[3] Zadeh, L.A.: Fuzzy sets, Inf. Control, Vol. 8, pp338–353, 1965.

[4] Pawlak, Z.: Rough Set- Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dorderecht, Boston, London (1991)

[5] Au, A., Chan, K.C.C., Wong, A.K.C., Wang, Y.: Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data, IEEE/ACM Transactions on Computational Biology and Bioinformatics,Vol.2, No.2, pp83-101, 2005.

[6] Wang, Y. Tetko, I.V., Hall, M.A., Frank, E., Facius, A., Mayer, K.F.X., Mewes, H.W.: Gene Selection from Microarray Data for Cancer Classification—A Machine Learning Approach, Computational Biology and Chemistry, Vol. 29, No., pp 37–46, 2005.

[7] Ding, C.: Analysis of Gene Expression Profiles: Class Discovery and Leaf Ordering, Proceedings of 6th Annual Conference on Research in Computational Molecular Biology, ACM Press, New York, pp127-136, 2002.

[8] Banerjee, M., Mitra, S., Banka, H.: Evolutionary rough feature selection in gene expression data systems, Man and Cybernetics, Part C: Applications and reviews 37:pp622-632, 2007.

[9] Wang, J., Waog, J.: Reduction Algorithms Based on Discernibly Matrix: The Ordered Attributes Method, Journal of Computer Science And Technology, Vo1.16, No.6, pp 489-504, 2002.

[10] Grzymala-Busse, J. W.: LERS-a system for learning from examples based on rough sets. In *Intelligent Decision Support*, Slowinski, R. (ed.) : Dordrecht: Kluwer Academic Publishers, pp3–18,1998.

[11] Hall, M.A.: Correlation-based feature selection for machine learning, Ph.D. Thesis. Department of Computer Science, University of Waikato (1999)

[12] Li, J., Liu, H., Downing, J.R., Yeoh, A.E., Wong, L.: Simple Rules Underlying Gene Expression Profiles of More Than Six Subtypes of Acute Lymphoblastic Leukemia (ALL) Patients, Bioinformatics, Vol. 19, No.1 pp71–78, 2003.

[13] Jensen, R., Shen, Q.: Semantics-preserving dimensionality reduction: rough and fuzzy-rough based approaches, *IEEE Trans. on Knowledge and Data Engineering*, vol. 16, no. 12 pp1457–1471, 2004.

[14] Jensen, R.: Combining rough and fuzzy sets for feature selection, Ph.D. Dissertation, School of Informatics , University of Edinburgh, Edinburgh (2004).

[15] Dash, S.: A Rule Induction Model Empowered by Fuzzy-Rough Particle Swarm Optimization Algorithm for Classification of Microarray Dataset, International Conference on Computational Intelligence in Data Mining (ICCIDM 2014), presented and published in Springer proceedings: Smart Innovation, System and Technology. (Accepted)

[16] Ding, C., Peng, H.C.: Minimum Redundancy Feature Selection from Microarray Gene Expression Data, Journal of Bioinformatics and Computational Biology, Vol.3, No.2 pp185-205, 2003.

[17] Sun Lijun, Miao Duoqian and Zhang Hongyun: Gene Selection with Rough Sets for Cancer Classification Fourth International Conference on Fuzzy Systems and knowledge Discovery, IEEE Computer Society (2007)

[18] Statnikov, A., Aliferis, C.F., Tsamardinos, I., Hardin, D., Levy, S.: A comprehensive evaluation of multi-category classification methods for microarray gene expression cancer diagnosis, Bioinformatics, Vol. 21, No. 5, pp 631–643, 2005.

[19] Akadi, A.E., Amine, A., Ouardighi, A.E., Aboutajdine, D.: Feature selection for Genomic data by combining filter and wrapper approaches, INFOCMP Journal of computer science, vol. 8, no. 4, pp28-36, 2009.