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

GS-RSAR : A Technique for Feature Selection of Microarray Data

Bichitrananda Patra[1]

[1]*KMBB College of Engineering &Technology, Bhubaneswar, Odisha, India*

Abstract- Gene expression profiles have great potential as a medical diagnostic tool since they represent the state of a cell at the molecular level. Available training data sets for classification of cancer types generally have a fairly small sample size compared to the number of genes involved. This fact poses an insurmountable problem to some classification methodologies due to training data limitations. Feature selection is considered a problem of global combinatorial optimization in machine learning, which reduces the number of features, removes irrelevant, noisy and redundant data, and results in acceptable classification accuracy. Hence, selecting relevant genes from the microarray data poses a formidable challenge to researchers due to the high-dimensionality of features, multi-class categories being involved, and the usually small sample size. To overcome this difficulty, a good selection method for genes relevant for sample classification is needed in order to improve prediction accuracy, and to avoid incomprehensibility due to the large number of genes investigated. In this paper, irrelevant genes are eliminated in two stages, employing correlation-based feature selection (CFS)as an evaluator and genetic search (GS)as a search technique at the first phase and in the second phase of elimination it an implementation of the Quick reduct algorithm of rough set attribute reduction(RSAR) and in third phase of elimination is (CFS+GS) combines Quick-Reduct algorithm and forms an integrated filter method. Since the data consist of a large number of redundant features, an initial redundancy reduction of the gene is done to enable faster convergence. Then Rough set theory is employed to generate reducts, which represent the minimal sets of non-redundant gene capable of discerning between all objects, in a multi-objective framework. The effectiveness of the proposed approach was verified on six different binary and multi-class microarray datasets using four different ANN classifier as LVQ1, LVQ2, OLVQ1 and SOM with 10 fold cross-validation method.

Keywords- feature selection, genetic search, quick- reduct, LVQ, SOM**.**

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 practice, filtering and classification algorithms are widely adopted to analyze gene expression data[1][3][4][5][7][12][14][15], in this paper, we focus on cancer classification using gene expression data, which is a hot topic in recent years and has received general attention by many biological and medical researchers. A reliable and precise classification of tumors based on gene expression data may lead to a more complete understanding of molecular variations among tumors, and hence, to better diagnosis and treatment strategies.

Feature selection also helps people to acquire better understanding about their data by telling them which are the important features and how they are related with each other[1][13][14][15].

Feature selection procedures output a list of relevant genes which may be experimentally analyzed by biologists. This method is often denoted as univariate feature selection (filter method), whose advantages are its simplicity and interpretability. However, because the interactions and correlations among the genes are omitted, filter method fails to remove redundant genes. The scores assigned to correlated genes are too similar, and none of the genes are strongly preferred over others. Redundancy among selected genes results in two problems. One problem is that redundant features in the selected subset reduce the comprehensive representation of target labels. The other one is that redundant genes increase the dimensionality of the selected gene set, which affect the mining performance on the small sample.

In many cases, sequence search is inclined to be trapped in local best solution. For the problem of optimization, Swarm Intelligence algorithms are effective methods, which involve algorithmic mechanisms similar to natural evolution and social behavior respectively. Genetic Search (GS) algorithm is one of the most common paradigms of such methods, which was proposed by Eberhart and Kennedy [2]. During the last ten years, GS gained increasing popularity due to its simplification and effectiveness.

Rough set theory (RST) has been used as a tool to discover data dependencies and to reduce the number of attributes contained in a dataset using the data alone, requiring no additional information [16],[17]. Over the past ten years, 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; all other attributes can be removed from the dataset with minimal information loss. A quick search of biological literatures shows that rough sets are still seldom used in bioinformatics. A major obstacle for using rough sets to deal with gene expression data may be the large scale of gene expression data and the comparatively slow computational speed of rough sets algorithms.

In order to find a satisfying gene set with minimum redundancy, we apply by combining correlation based feature selection and Genetic Search (CFS-GS) with supervised quick reduct algorithm (CFS-GS-QR). The fitness function of RSAR and GS explicitly measure the feature relevant and feature redundancy simultaneously. CFS-GS finds a compact feature set with great predictive ability due to the high efficiency of GS algorithm. Then supervised quick reduct (QR) algorithm is employed that enables CFS-GS to find reducts, which represent the minimal sets of non-redundant features capable of discerning between all objects. The effectiveness of the algorithm is demonstrated on six benchmark binary and multi-class cancer datasets viz. Colon, Leukemia, Lung, Prostrate, Leukemia_GEMS and Lung_GEMS cancer.

This paper follows Section 2 briefly introduces Rough set theory and GS algorithm Feature Selection and novel algorithm (GS-RSAR) is explained in briefly in section 3. Then, data sets and experiment settings are described in section 4. We show the results and discussions in section 5. Finally, conclusions are given in section 6.

II. ROUGH SET THEORY AND GENETIC SEARCH

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 $P \subseteq A$, there is an associated equivalence relation:

The partition of U generated by IND(P) is denoted U/P . If (x,y) $\in$ IND(P) , then x and y are indiscernible by attributes from P . The equivalence classes of the P-indiscernibility relation are

denoted by $[x]_p$. Let $X \subseteq U$, the P-lower approximation $\underline{P}X$ and P-upper approximation $\overline{P}X$ of set X can be defined as:

$$\underline{P}X = \{x \in U \mid [x]\, p \subseteq X\}$$
$$\overline{P}X = \{x \in U \mid [x]p \cap X \neq \phi\}$$

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

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

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 $P \underset{k}{\Rightarrow} Q$

$$k = \gamma_P(Q) = \frac{|\mathrm{POS}_P(Q)|}{|U|}$$

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 :

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

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

$$\mathrm{Red(C)}_{min} = \{R \epsilon \mathrm{Red} | \forall R` \epsilon \mathrm{Red}, |\mathrm{R}| \leq |\mathrm{R`}|\}$$

Reducts [9] obtained in a decision table usually is more than one, generally the reduct with the fewest attributes is optimal. Obtaing all reducts or minimal reducts of a decision table is a NP-hard problem, thus heuristic knowledge deriving from the dependency relationship is mainly used to assist the attribute reduction [10] .

Genetic Search Feature Selection

Standard genetic operators, such as crossover and mutation, are applied without modification. Crossover is used to swap the genetic material of chromosomes between selected couples to produce new offspring's that are capable of preserving the characteristics of the parent chromosomes well. Many kinds of crossover procedures have been tried in GSs to date. In this study, a 2-point crossover operator was used, which chose two cutting points at random and alternately copied single segments out of each parent. If a mutation was present, either one of the offspring was mutated and its binary representation changed from 1 to 0, or from 0 to 1 after the crossover operator is applied. If the mutated chromosome was superior to both parents, it replaced the worst chromosome of the parents; otherwise, the most inferior chromosome in the population was replaced. The GS was configured to contain 30 populations and was run for 100 generations in each configuration. The crossover and mutation rates were 0.8 and 0.1, respectively.

III. PROPOSED NOVEL ALGORITHM: GS-RSAR

Genetic Algorithm (GS) was first proposed by Holland in 1975 (Holland, 1975). GS is inspired by Darwin's theory of evolution [6]. In this chapter, the above two different feature selection models for microarray data classification were combined to select relevant genes The features selected during the first-stage were used for feature selection by the genetic algorithm. The GS population is

initialized randomly, with each chromosome in the population coded to a binary string. The chromosome length represents the number of the features. The bit value {1} represents a selected feature, whereas the bit value {0} represents a non-selected feature. The predictive accuracy of a 1-NN determined by the LOOCV method was used to measure the fitness of an individual. For example, when a 9-dimensional data set (n = 9) is analyzed, any number of features smaller than n can be selected. When the adaptive value is calculated, these five features in each data set represent the data dimension and are evaluated by the 1-NN method. The fitness value for 1-NN evolves according to the LOOCV method for all data sets. In the LOOCV method, a single observation from the original sample is selected as the validation data, and the remaining observations as the training data. This is repeated so that each observation in the sample is used once as the validation data.

In the second-stage, RSAR, a filter method, was used to reduct from the subset generated from genetic search. It has two parameters, conditional attribute and decision attribute and its evaluation of degree of dependency value leads to the decision attribute[8]. 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[11]. According to the algorithm, the dependency of each attribute is calculated and the best candidate is chosen. The performance of supervised algorithm will be GS-RSAR examined in our experiments.

Step 1: Encode each gene in a chromosome is a bit, which takes a value of either 1 or 0.

Step 2: Randomly select a pool of chromosomes from the entire population. In this study, we initial the population size was set to 100 chromosomes.

Step 3: Evaluate the fitness value (prediction accuracy) of each chromosome in the pool and update the time available for searching.

Step 4: Calculate cumulative selection probabilities in the following sequence.

First, arrange chromosomes in descending order of their fitness scores (i.e. Prediction accuracies).

Second, rank the chromosomes on the basis of their fitness. Assign rank 1 to the lowest fitness score, 2 to the next and so on.

Third, estimate the probability of selection for each chromosome and then calculate the cumulative selection probability of each chromosome.

Step 5: Each couple creates two offspring by crossover and then the parents die.

Step 6: Perform mutation on the offspring. If the mutation rate is very high, the GS descends into a random search. Conversely, too low a mutation rate implies too little exploration in the search space.

Step 7: Repeat steps 3 to 7 until the stopping criterion is satisfied.

Step 8: $R \leftarrow \{\ \}$

Step 9: do

Step 10: $T \leftarrow R$

Step 11: $\forall x \in (C - R)$

Step 12: if $\gamma_{R \cup \{x\}}(D) > \gamma_T(D)$

Step 13: $T \leftarrow R \cup \{x\}$

Step 14: $R \leftarrow T$

Step 15: until $\gamma_R(D) == \gamma_C(D)$

Step 16: return R

IV. DATA SETS

Six used microarray gene expression data sets are chosen for our experiments: Colon tumor, ALL-AML Leukemia, Lung cancer, Prostate_Tumor_GEMS, Brain_Tumor_GEMS and Leukemia_GEMS. The first three data is taken from http://sdmc.lit.org.sg/GEDatasets/Datasets.html and other data is taken from http://www.gems-system.org. Table 1. summarize these datasets. We conducted the experiments on these six data sets by applying Rough Set Attribute Reduction (RSAR) Subset Evaluation method for feature reduction and Learning Vector Quantisation (LVQ) - LVQ1, LVQ3, optimized-learning-rate LVQ1 (OLVQ1), and The Self-Organizing Map (SOM) [19] for neural classification of the reduced datasets. We used Weka, a well-known comprehensive toolset for machine learning and data mining [18] as our main experimental platform. We evaluated the performance of feature reduction in Weka environment with four classifiers, using 10-fold Cross Validation.

Table 1. Dataset Information

Dataset	# classes	# instances	# attributes
Colon tumor (Train/	2	40 / 22	2000
Leukemia (Train/	2	47 / 25	7129
Lung cancer (Train/	2	109 / 72	12533
Prostate Tumor	2	61 / 41	10509
Leukemia_GEMS	3	43 / 29	11225
Lung_GEMS (Train/	5	43 / 29	11225

V. EXPERIMENTAL RESULTS

The classifier tool WEKA [20] is 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 WEKA is a well-known package of data mining tools which provides a variety of known, well maintained classification and filter algorithms. This allows us to do experiments with several kinds of classifiers quickly and easily. The tool is used to perform benchmark experiment. Four classifier learners were employed for the classification of the data, LVQ1, LVQ2, LVQ3 and SOM[19].

Table 2 consists the number of genes attribute of above given train datasets after the GS, RSAR and GS-RSAR feature selection method.

Table 2. No.of gene expression after feature selection

Dataset	# of Genes	# of Genes after		
		GS	RSAR	GS-
Colon tumor	2000	518	9	8
Leukemia	7129	2809	3	2
Lung cancer	12533	4852	13	2
Prostate Tumor	10509	5145	14	13
Leukemia_GEMS	11225	4290	14	14
Lung_GEMS	11225	2401	54	25

Many researchers are currently studying how to select genes effectively before using a classification method to increase the predictive accuracy. In general, gene selection is based on two aspects: one is to obtain a set of genes that have similar functions and a close relationship, the other is to find the smallest set of genes that can provide meaningful diagnostic information for disease prediction without diminishing accuracy. Feature selection uses relatively fewer features since only selective features need to be used. This does not affect the predictive accuracy in a negative way; on the contrary, predictive accuracy can even be improved.

To evaluate the performance of the proposed GS-RSAR, the experimental results obtained from the proposed GS-RSAR were compared with the results of GS and RSAR methods reported in the literature.

We compare the performance accuracies of different classifiers, LVQ1, LVQ3, OLVQ1 supervised classifiers and one unsupervised classifier SOM on four binary and two multi class datasets using 10-fold Cross Validation (CV) without using feature selection algorithms. The results [19] of the 10-fold CV accuracy for the classifiers are shown in table 3. After feature selection, the selected feature subsets were evaluated using the above classification algorithms using 10-fold CV method are shown in Table 4.

The LVQ was configured to change Intialisation mode as K-Nearest Neighbour Even, learning function as static and was run for 1000 training iterations using voting. The code book vector and learning rates were 30 and 0.2 respectively.

Table 3. Classification accuracy by using four ANN classifier before subset evaluation

Datasets	Lvq1	Lvq3	Olvq1	SOM
Colon	85	80	90	47.5
Leukemia	95.7447	95.7447	87.234	57.4468
Lung	98.1651	96.3303	95.4128	75.2294
Prostate Tumor	75.4098	81.9672	80.3279	68.8525
Leukemia_GEMS	90.6977	93.0233	93.0233	62.7907
Lung_GEMS	90.9836	89.3443	88.5246	84.4262
Average	89.33348	89.40163	89.0871	66.04093

Table 4. Classification accuracy by using four ANN classifier after RSAR subset evaluation .

Datasets	Lvq1	Lvq3	Olvq1	SOM
Colon	87.5	87.5	87.5	67.5
Leukemia	95.7447	95.7447	95.7447	78.7234
Lung	97.2477	97.2477	97.2477	88.9808
Prostate Tumor	93.4426	91.8033	91.8033	88.5246
Leukemia2_GEMS	90.6977	90.6977	90.6977	81.3953
Lung_GEMS	72.1331	69.6721	68.8525	68.0328
Average	89.46097	88.77758	88.64098	78.85948

Table 5. *Classification accuracy by using four ANN classifier after GS subset evaluation .*

Datasets	Lvq1	Lvq3	Olvq1	SOM
Colon	95	80	90	70
Leukemia	95.7447	95.7447	93.617	76.5957
Lung	96.3303	95.4128	94.4954	82.5688
Prostate Tumor	73.7705	83.6066	80.3279	55.7377
Leukemia_GEMS	95.3488	90.6977	90.6977	55.814
Lung_GEMS	90.9836	90.1639	89.3443	68.8525
Average	91.19632	89.27095	89.74705	68.26145

Table 6. *Classification accuracy by using four ANN classifier after GS-RSAR subset evaluation.*

Datasets	Lvq1	Lvq3	Olvq1	SOM
Colon	80	80	80	77.5
Leukemia	85.1064	82.9787	96.5957	76.5957
Lung	99.0826	99.0826	99.0826	95.4128
Prostate Tumor	72.1311	80.3279	77.0492	60.6557
Leukemia_GEMS	83.7209	79.0689	86.0645	81.3953
Lung_GEMS	80.3279	81.9672	77.0492	68.0328
Average	83.39482	83.90422	85.97353	76.59872

Table 2 shows that the number of necessarily selected features could be greatly reduced by the proposed method. Table 5 compares experimental results obtained by Table 3 and Table 4. i.e. GS, RSAR feature selection methods and the proposed GS-RSAR method. Various methods were compared to the proposed method. These include: supervised LVQ : (1) LVQ1, (2) LVQ3, (3) OLVQ1. The unsupervised method included: SOM. We compared the classification accuracy with the accuracy achieved LVQ and SOM methods because the LVQ classifiers usually results in higher accuracy than SOM classifier.

The experimental results show that the accuracy of microarray data which had featured selection by GS and RSAR implemented were better than without feature selection datasets. Comparing binary class and multi class datasets, the accuracy of the binary data sets was better than the multi class data sets. But the proposed GS-RSAR was not given the better result.

Again, we can observe from Table 3, Table 4 and Table 5, that the proposed method effectively increases classification accuracy and selects a small number of feature subsets. From Table 2 It is observed that during the GS-RSAR subset evaluation of the proposed method returns very small sets of genes compared to alternative variable selection methods, while retaining predictive performance. Our method of gene selection will not return sets of genes that are highly correlated, because they are redundant. This method will be useful when considering the design of diagnostic tools, where having a small set of probes is often desirable and to help understand the results from other gene selection approaches that return many genes, so as to understand which ones of those genes have the largest signal to noise ratio and could be used as surrogates for complex processes involving many correlated genes.

The average highest classification accuracies of RSAR/LVQ1, RSAR/LVQ3, RSAR/OLVQ1, RSAR/SOM, GS/LVQ1, GS/LVQ3, GS/OLVQ1, GS/SOM were 91.7943, 90.44425, 90.30765, 78.85948, 91.16932, 89.27095, 89.74705, 68.26145 respectively. Based on the comparison of results produced by SOM and LVQ algorithms on the microarray gene expression datasets, LVQ produced better results than SOM, and out of the three LVQ algorithms LVQ1 was the best.

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 the other data set, the reason maybe because the scale of lung data set is larger. In theory, the more information available about the problem, the more likely for rough sets method to finding informative features. If possible we can get more accurately predicted result through constructing as large data set as possible to train a rough set model.

VI. CONCLUSION

We conducted an extensive survey in the area of building classification models from microarray data with various ANN classification algorithms. Experimental results show that in most cases, the LVQ algorithms delivered classification accuracies equivalent to or better than those on the same data sets reported by other studies. GS and RSAR subset valuation method, which is proving to be an

appropriate feature selection method, the learning algorithms are capable of building classification models with high predictive accuracies of microarray data. The classification accuracy of lung cancer data set is higher than the other data set in proposed GS-RSAR feature selection method. As the study shows that the feature reduction scheme improves classification accuracies, one question immediately arises: will there be better schemes for the feature selection process for building ANN classification models? Since the number of instances in the studied microarray data is small and the performances of many classification algorithms are sensitive to the number of training data, another interesting question is raised: when comparing predictive performances of various classification algorithms on microarray data, what is the impact of adopting different feature selection methodologies.

REFERENCES

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

[2]. H. Liu, E. Dougherty, J. Dy, K. Torkkola, E. Tuv, H. Peng, C. Ding, F. Long, M. Berens, L. Parsons, Z Zhao, L. Yu, and G. Forman, "Evolving feature selection," *IEEE INTELLIGENT SYSTEMS*, vol. 20, no. 6, pp. 64–76, 2005.

[3]. J. Li, H. Liu, J.R. Downing, A.E. Yeoh, and L. Wong,"Simple Rules Underlying Gene Expression Profiles of More Than Six Subtypes of Acute Lymphoblastic Leukaemia (ALL) Patients", Bioinformatics, Vol. 19, No.1, 2003, pp. 71–78.

[4]. Au, K.C.C. Chan, A.K.C. Wong, and Y. Wang,"Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data", IEEE/ACM Transactions on Computational Biology and Bioinformatics,Vol.2, No.2, June 2005, pp. 83-101.

[5]. F.D. Smet, NLMM. Pochet, K. Engelen, T.V. Gorp, P.V. Hummelen, K. Marchal, F. Amant, D. Timmerman, B.D.Moor, and I. Vergote, "Predicting the Clinical Behaviour of Ovarian Cancer from Gene Expression Profiles", International Journal of Gynaecological Cancer, Vol.16, No.s1, Feb.2006, pp.147–151.

[6]. Nguyen, D., Rocke, D. M., Tumor classification by partial least squares using microarray gene expressio data. Bioinformatics, 2002, 18, 39–50.

[7]. Y. Wang, I.V. Tetko, M.A. Hall, E. Frank, A. Facius, K.F.X. Mayer, and H.W. Mewes, "Gene Selection from Microarray Data for Cancer Classification—A Machine Learning Approach", Computational Biology and Chemistry, Vol. 29, No.1, Feb 2005, pp.37–46.

[8]. J.R. Quinlan, "Induction of Decession Trees : Machine Learning",vol.1,pp.81-106,1986.

[9]. Z. Pawlak, "Rough Set- Theoretical Aspects of Reasoning about Data", Kluwer Academic Publishers, Dorderecht, Boston, London, 1991.

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

[11]. J. R. Quinlan, "C4.5: Programs for Machine Learning", San Mateo,CA, Morgan Kaufmann Publishers,1993.

[12]. J. J. Valdes, A.J. Barton, **"**Gene Discovery in Leukaemia Revisited: A Computational Intelligence Perspective", Proceedings of the 17th International Conference on Industrial & Engineering Applications of Artificial International Conference & Expert Systems, Springer Verlag, 2004, pp.118-127.

[13]. Ding, H.C. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data", Journal of Bioinformatics and Computational Biology, Vol.3, No.2 , Apr. 2003, pp.185-205.

[14]. L.P. Wang, C. Feng, and X. Xie, "Accurate Cancer Classification Using Expressions of Very Few Genes", IEEE /ACM Transactions on Computational Biology and Bioinformatics, Vol.4, No.1, 2007, pp.40-53.

[15]. S.Mitra, Y.Hayashi, "Bioinformatics with Soft Computing", IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews, Vol. 36, No.5, 2006, pp.616-635.

[16]. J. W. Grzymala-Busse, "LERS-a system for learning from examples based on rough sets," in *Intelligent Decision Support*, R. Slowinski, Ed. 1992, Dordrecht: Kluwer Academic Publishers, pp. 3–18.

[17]. P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 1–13, 2002.

[18]. Ian H. Witten and Eibe Frank. "*Data Mining: Practical Machine Learning Tools and Techniques*." Morgan Kaufmann, San Francisco, 2 edition, 2005.

[19]. B.N. Patra, S. Dash, B.K. Tripath, "*Neural Techniques for Improving the Classification Accuracy of Microarray Data Set using Rough Set Feature Selection Method",* International Journal of Computer Trends and Technology, Volume4, Issue3, 2013.

[20]. L. Blake and C. J. Merz, UCI Repository of machine learning databases. [Online]. Available: http://www.ics.uci.edu/~mlearn/.