Asian Journal of Information Technology 15 (2): 199-210, 2016
ISSN: 1682-3915

Modified Fuzzy Rough Quick Reduct Algorithm for Feature Selection in Cancer Microarray Data

[1]C. Arunkumar and [2]S. Ramakrishnan
[1]Department of Computer Science and Engineering, Amrita School of Engineering, Coimbatore, India
[2]Department of Information Technology, Mahalingam College of Engineering and Technology, 642003 Pollachi, Tamil Nadu, India

Abstract: This study proposes a novel method that employs correlation based filter for dimensionality reduction followed by fuzzy rough quick reduct for feature selection on a particle swarm optimization search space. The first phase removed the redundant genes using correlation coefficient filter on a particle swarm optimization search space. The second phase produced a fuzzy rough quick reduct that would be used for classification. The genes obtained after feature selection are subjected to classification using traditional classifiers. It has been determined that the proposed method contributes to reduction in the total number of genes and improvement in the classifier accuracy compared to gene selection and classification using correlation coefficient and traditional fuzzy rough quick reduct algorithm. This approach also reduces the number of misclassifications that might occur in other approaches.

Key words: Feature selection, fuzzy rough set, correlation coefficient, particle swarm optimization, quick reduct

INTRODUCTION

Microarray data has stimulated a new arena of research in the fields of machine learning and bioinformatics in the last 20 year. The gene expression information obtained from microarray samples play a key role in disease diagnosis and treatment especially in the area of oncology that could identify and treat a variety of tumors. Microarray cancer gene expression data is composed of very small samples (usually>200) and thousands of gene expression levels (ranging from 7000-20000). Another essential component is the validation of the data. Further, microarray gene expression data becomes more exciting due to the presence of noise and outliers (Canedo *et al*., 2014). A single experiment could be carried out to monitor and measure the gene expression activation levels by the usage of microarray technology. This technique is widely adopted in the analysis and diagnosis of a large number of diseases. Large amount of data useful for solving many biological problems can be generated by a technique called microarray. Microarray is a technique which measures the level of activity of thousands of genes concurrently. If the gene is overexpressed then there will be too much protein which gives the conclusion that the particular gene is abnormal. Even much smaller changes can be detected by microarrays compared to karyotypes. The domain where microarray is used in the recent years is in disease classification. Gene expression data is data rich and information poor. Public microarray databases include NCBI, Genbank, Array Express, Gene Expression Omnibus and Stanford Microarray (Liu, 2008). Microarray platforms include Agilent, Affymetrix and Illumina Bead Array (Bennet *et al*., 2014). The microarray dataset that is used in our study is of the following format as in Table 1 where: G1, G2, G3, ..., Gn indicates the gene ID, 1, 2, ..., N indicates the instances which represents the data of each sample and the nth column indicates the class and the numerical values represent the gene expression levels. In this case, all values in the sample table lie between 0 and 1 which means that the data is normalized.

Table 1: Gene expression format used for our study

Gene ID	G_1	G_2	G_3	...	...	G_n	Class
1	0.2	0.1	0.2	0.3	0.7	0.9	Normal
2	0.3	0.1	0.5	0.4	0.1	0.7	Tumor
.	0.6	0.6	0.1	0.8	0.2	0.4	Tumor
N	0.8	0.7	0.8	0.9	0.1	0.6	Normal

The key problem in the area of pattern recognition is feature selection. The outcome of successful feature selection has multi-fold advantages like reduction in computational complexity and cost if proper dimensionality reduction techniques are applied,

Corresponding Author: C. Arunkumar, Department of Computer Science and Engineering, Amrita School of Engineering, Coimbatore, India

noise reduction that aids in increase of classifier accuracy and obtaining interpretable features that help in efficient disease diagnosis and treatment. Few genetic alterations occur biologically that correspond to the malignant tumor in cells. Efficient process lies in finding those regions of interest that could help in investigating the cause of the disease (Maulik and Chakraborty, 2014; Chakraborty and Maulik, 2014). The current focus is to increase the classifier accuracy and to perform efficient feature selection. The main aim of research in the area of classification accuracy involves prediction of the class membership of the data, production of the correct label for the training data and predicting the labels of unknown data with higher degree of accuracy (Yang *et al*., 2008). The small training and testing data available and their higher dimensionality increases the difficulty level of the classification task. Significant correlation would be exhibited by a very few genes of a particular phenotype but requires investigation of thousands of gene samples. So, feature selection is a very crucial procedure to understand and analyze the gene expression profiles and hence aid in achieving higher classification accuracy. The classification accuracy of unknown samples plays a crucial role in a medical diagnosis system (Ghorai *et al*., 2011).

Feature selection methods fall into three broad categories in data mining namely the filter, wrapper and hybrid approaches (Yang *et al*., 2008). The process of classification is performed after filtering in the filter model approach (Dash *et al*., 2012). The weight value for each feature is computed and higher values are chosen to represent the reduced feature subset. The statistical properties of the data contribute majorly in the relevant feature selection process using the filter model. This approach reduces the dimensionality of datasets independent of the learning algorithm. The interaction between features is not considered in the filter approach and this is one major disadvantage of this model.

The wrapper approach resembles an optimal algorithm that searches for an optimal solution in a given dimensional space (Liu, 2008). The wrapper approach utilizes a given learning algorithm to evaluate the candidate feature subsets. Hence, the feature selection process is tied to the learning algorithm in a wrapper model. Three main issues in a wrapper model makes it challenging. They are search operation on a high dimensional space called the NP complete problem, uncertain assessments that make the choice of feature configuration difficult and the high dimensionality of a given problem that makes the selection of a feature subset complex (Bontempi, 2007).

The drawbacks in the filter and wrapper approaches could be eliminated by using hybrid approaches. The hybrid model makes use of a combined filter-wrapper model. It uses the simplicity nature of the filter model at the initial gene selection level in combination with the optimized wrapper approach that increases the classification accuracy at the final stage. The hybrid model works in two stages. In the first stage, the filter model is applied. The filter eliminates irrelevant and redundant genes from the original dataset. The resultant data applied after the filter model contains much lesser number of genes. In the second stage, the wrapper is applied on the filtered dataset and the training accuracy is optimized. This approach brings the hybrid model to an acceptable level of performance and satisfaction. A hybrid approach combining correlation based feature selection and linear forward selection was performed on three microarray gene expression datasets. Later traditional classifiers were used to evaluate their performance. The hybrid approach selected lesser number of genes compared to filter based approaches. Also, the hybrid approach of feature selection performed comparatively better or equal to the filter based feature selection approaches (Arunkumar and Ramakrishnan, 2015). The embedded approach that associates itself with a specific learning algorithm seeks to subsume feature selection as part of the model building process. On the other hand, the goal of the hybrid approach is to take advantage of both the filter and the wrapper approaches (Leung and Hung, 2010).

Fuzzy rough set theory evolved from the rough set theory is considered as one of the important tools in granular computing. This concept attracts wider attention from several domains that include machine learning, granular computing and uncertainty reasoning. The concepts of roughness and fuzziness are encapsulated into a single model. The two elemental modules of human reasoning that would be imitated by the fuzzy rough set and cognition are fuzzy information granulation and approximate reasoning. The fuzzy concepts of the universe are formed by using the attributes and also describe other objects. This theory granulates the universe of discourse into a set of fuzzy concepts based on fuzzy relations and then approximates arbitrary fuzzy sets with these fuzzy concepts (Hu *et al*., 2012).

The inconsistency between the decision labels and condition attributes is addressed by using the concept of rough set theory. In rough set theory, the universe of discourse is partitioned by the decision labels and condition attributes. Samples that possess the same description are aggregated after the process of partition has been performed. If two samples with the same description belong to two different decision labels, there arises an inconsistency among the partition that

comprises of the decision labels and condition attributes. Deletion of a condition attribute would increase the level of inconsistency. The main aim of rough set theory attribute reduction is to maintain the original inconsistency. This could be achieved by deleting superfluous attributes and obtain a reduced set of condition attributes by performing discerning between the two samples. The concept of attribute reduction and feature selection in fuzzy rough set is a purely structural approach that depends only on the dataset. It does not require any other additional knowledge. The discernible ability of condition attributes related to decision labels is highlighted and this serves as a key differentiator between the rough set based feature selection and other traditional approaches that includes the filter and wrapper models (Chen and Yang, 2014).

Rough set theory proposes a number of methods for feature selection using heuristic based techniques. Some methods are stated as under: a feature subset that could be distinguished by any two objects by using the concept of discernibility is discussed by Skowron (1995). Attribute reduction using positive region is discussed by Grzymala-Busse in 1991. The same kind of approach with target decision unchanged is discussed by Hu and Cercone (1995). The concept of using information entropy to search 'reduct' in a rough set model is discussed by Slezak in 2000. The concept is expanded to approximate reduct that could be used for a number of feature reduction methods is discussed by Ziarko (1993) and Dai and Xu (2013).

A typical classification task involves binary classification of the given problem as either "normal" or "cancerous". In certain cases, it might also involve classifying a multi-class problem that involves classification among the different types of cancer. The presence of binary and multi-class problems in microarray gene expression data is of serious concern to researchers worldwide. The main challenge is the existence of high dimensional data with a very small sample size. This "large p, small n" problem is called the curse of dimensionality. Many dimensionality reduction algorithms have been developed to avoid this phenomenon. So, it is absolutely essential to build a robust model that could perform the process of feature selection and classification. At this juncture, we propose a modified quick reduct algorithm for feature selection combining the correlation based filter for dimensionality reduction with the fuzzy rough quick reduct algorithm on a particle swarm optimization search space for microarray cancer gene expression data. This method has the capability to reduce the number of genes and also increase the classifier accuracy.

Table 2: Description of the dtaset used for the study

Dataset name	No. of genes in raw dataset
ALL/AML	7,129
Lung Cancer	12,533
Ovarian Cancer	15,154

This study compares the proposed feature selection approach of microarray gene expression data against traditional approaches. Six classification algorithms are used to compare and evaluate the classifier accuracy of the proposed method with correlation based filter and fuzzy rough quick reduct algorithm. The proposed method shows better results in terms of the number of feature gene subsets selected and classifier accuracy compared to traditional methods of feature selection like correlation based filter and fuzzy rough quick reduct.

Dataset description: We used three binary category cancer-related human gene expression data sets which were downloaded from the Kent ridge biomedical repository to evaluate the performance of the proposed method. The data format is shown in Table 2; it includes the data set name, the number of genes, number of training and testing samples.

Acute lymphoblastic leukemia is an acute form of leukemia caused by overproduction and accumulation of immature white blood cells called as lymphoblasts. Acute myeloid leukemia is caused by the rapid and abnormal growth of white blood cells that affect the bone marrow. It also interferes with the production of normal blood cells. The dataset consists of 72 samples out of which 47 are ALL and 25 are AML.

Lung Cancer or pulmonary carcinoma derived from epithelial cells is caused by malignant uncontrolled cell growth in the lung tissues. This type of cancer is harmful as it has higher probability to spread to other parts of the body if left untreated. The dataset consists of 181 tissue samples out of which 150 belong to ADCA and 31 are MPM.

Ovarian cancer is a type of cancer that begins in the ovaries. This develops in women who have higher risk of development due to family or personal history. The dataset consists of 253 samples out of which 162 are ovarian cancer and 91 are controls (normal).

MATERIALS AND METHODS

Proposed approach to feature selection: The proposed method combines the concept of correlation coefficient and fuzzy rough quick reduct algorithm on a particle swarm optimization search space. Pearson's coefficient indicates the amount of correlation that exists between

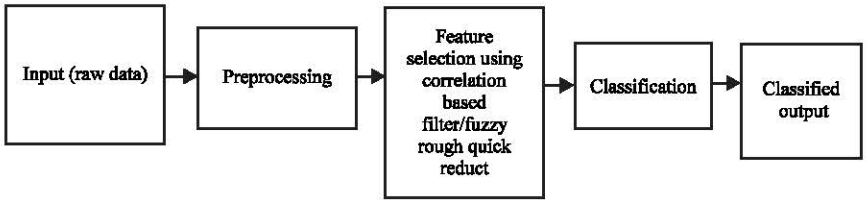

Fig. 1: Block diagram of the traditional approach to feature selection

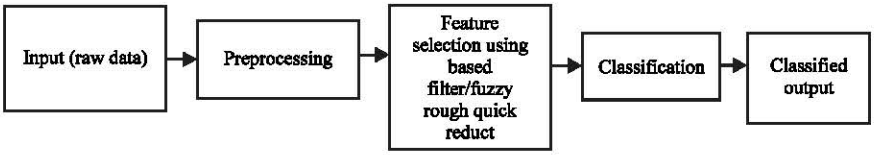

Fig. 2: Block diagram of the proposed approach to feature selection

each feature and the decision class in the dataset. The basic idea behind Pearson's coefficient is to select the features that reveal very important information about the different classes. Ideally, such features are highly discriminative and occur in a single class. Each feature in the dataset produces a correlation coefficient value that would be used to select or discard a particular feature (gene). Hence, it becomes necessary to define a suitable threshold value. The feature selection process is done if the Pearsons coefficient value is higher than the defined threshold. The block diagram for the traditional and proposed approach is given:

The block diagram shown in Fig. 1 shows the traditional approach to feature selection. In the first step, the raw gene expression dataset is taken as input. Normalization is done as a part of preprocessing. The feature selection process uses the concept of correlation coefficient and fuzzy rough quick reduct algorithm to select features and the classification is performed using traditional classifiers.

The block diagram shown in Fig. 2 shows the proposed approach to feature selection. In the first step, the raw gene expression dataset is taken as input. Normalization is done as a part of preprocessing. The feature selection process uses the modified Fuzzy Rough Quick Reduct algorithm for feature selection. The concepts of correlation coefficient on a particle swarm optimization search space coupled with Fuzzy Rough Quick Reduct algorithm to eliminate the redundant features and select only the essential ones is employed. Later classification is performed using traditional classifiers and the results are analyzed in further sections.

Preprocessing: The raw dataset obtained by experiments conducted on cancer microarray gene expression data consists of gene expression levels at various ranges. Normalization is performed to fit attribute data into a specific range, say [0,1]. Our datasets are normalized using min-max normalization. Min-max normalization transforms a gene expression value A-B to fit in the range [C,D]. The formula for min-max normalization is given in Eq. 1:

$$B = \left(\frac{\left(A - \text{minimum value of A} \right)}{\left(\text{maximum value of A} - \text{minimum value of A} \right)} \right) \times (D - C) + C \quad (1)$$

Using in Eq. 1, all our raw datasets are normalized to the range (0,1) (C,D) for further processing.

Correlation based feature selection on a particle swarm optimization search strategy: The correlation coefficient is computed by using a correlation based heuristic evaluation function. It uses a multivariate approach which considers the interaction among features (Fu and Youn, 2003; Hall, 1999; Lazar *et al* ., 2012). Subsets of attributes are evaluated using the identification ability of each of the attributes. Outliers and noise makes correlation coefficient highly sensitive (Costa *et al*., 2011). Statisticians use correlation to measure the degree of linear relationship that exists between Genes (Yang *et al*., 2008). Formula for calculating Pearson correlation between features x_i and y_i is given in Eq. 2:

$$\text{Correlation} = \sum \left(\frac{x_i - \text{mean}(x_i) \times y_i - \text{mean}(y_i)}{n \times SD(x_i) \times SD(y_i)} \right) \quad (2)$$

The correlation coefficient is computed and genes that exhibit low inter-correlation are selected and used to study different types of cancer (Kumar and Ramakrishnan,

2014). The computed correlation value lies in the range (0,1). Values >0.5 exhibits high degree of correlation and those with values in the range (0.3, 0.5) are said to have low degree of correlation. The normalized values obtained are fed as input to this feature selection phase. Pearsons coefficient of individual attributes are found out and attributes having values higher than a defined threshold are selected and fed to the different classifier algorithms and the results are analyzed. Particle Swarm Optimization or PSO is stochastic in nature with bio inspired behavior. It is an evolutionary algorithm with swarm intelligence embedded into it that works on the key concept of sharing of information. It is useful to solve a lot of engineering problems that uses several variables. This algorithm looks for food and allows the associated particles to get their benefits based on previous experience and earlier discoveries. A candidate solution is obtained for each of the particles that flies and passes through the search space (Arunkumar and Ramakrishnan, 2015).

Fuzzy Rough Quick Reduct algorithm: Let the full feature set obtained after feature selection phase-1 using Information gain filter is A. Consider the features obtained after fuzzy rough set attribute reduction is defined by a subset R. In terms of the dependency function, $\gamma(R)$ and $\gamma(A)$ would be identical if the dataset is consistent. If the dataset is consistent, Eq. 3 holds true:

$$\gamma(R) = \gamma(A) = 1 \tag{3}$$

However, in the case of fuzzy rough set based attribute reduction, the above equation may not hold true. This is because of the uncertainty encountered when features belong to many fuzzy equivalence classes that result in a reduced total dependency. In order to overcome this uncertainty, a dependency function γ^1 is defined which would aid in choosing features to add to the current reduct set. The reduct algorithm terminates when the addition of any remaining feature does not increase the dependency. The algorithm for Modified Fuzzy Rough Quick Reduct is as follows.

Modified fuzzy rough quick reduct algorithm: Input: Let R represent the raw dataset obtained after the process of normalization. R is subjected to dimensionality reduction using correlation coefficient filter on a particle swarm optimization search space which produces A, the set of all conditional features and B is the set of decision features (nth column in the dataset that determines the type of disease).

Let P represent the Pearsons coefficient and R denotes the raw dataset. The Pearsons correlation coefficient is determined using in Eq. 4 on a particle swarm optimization search space:

$$\text{Correlation} = \sum\left(\frac{x_i - \text{mean}(x_i) \times y_i - \text{mean}(y_i)}{n \times SD(x_i) \times SD(y_i)}\right) \tag{4}$$

- P←R
- Initialize x_i, v_i
- $p_i = x_i$
- Compute $p_i(\Delta t+1)$ using (Eq. 5)
- Compute gbest(Δt) using (Eq. 6)
- Compute updated velocity $v_{i,j}(\Delta t+1)$ using (Eq. 7)
- Compute updated position $x_i(\Delta t+1)$ using (Eq. 8)
- C<{ }; $\gamma^1_{best} = 0$; $\gamma^1_{prev} = 0$
- While $\gamma^1_{best} \neq \gamma^1_{prev}$
- T←C
- $\gamma^1_{prev} = \gamma^1_{best}$
- Foreach x (A-C)
- If $\gamma^1_{C\,U\,(x)}(B) > \gamma^1_T(B)$
- T←C U { x }
- $\gamma^1_{best} = \gamma^1_T(B)$
- C←T
- Return C

The particle's position can be biased by two factors namely the best position visited by the current particle and that of its neighboring particle. It is said to have the world's best particle if the neighborhood is a swarm of particles. The kind of optimization problem determines the fitness function which would yield the global optimum (Kar *et al*., 2015). Each particle in the swarm is represented by the following uniqueness:

- x_i = Current position of the ith particle
- v_i = Current velocity of the ith particle
- p_i = Best previous position of the ith particle
- gbest = global best particle in its neighborhood

The personal best position of particle i is the best position experienced by the particle so far. If f is the objective function, the personal best of a particle, at time step Δt is calculated as:

$$\begin{Bmatrix} p_i(\Delta t+1) = & p_i(\Delta t) & \text{if } f(x_i(\Delta t+1)) >= f(p_i(\Delta t)) \\ & x_i(\Delta t+1) & \text{if } f(x_i(\Delta t+1) < f(p_i(\Delta t)) \end{Bmatrix} \tag{5}$$

If g best denotes the global best particle, it is given as:

$$\begin{aligned} &gbest(\Delta t)\ \varepsilon\ \{p_0, p_1, \ldots p_s\} = \\ &\min\{f(p_0(\Delta t), f(p_1(\Delta t), \ldots f(p_s(\Delta t)) \end{aligned} \tag{6}$$

where, s is the size of the entire swarm. The velocity of the particle i is updated by:

$$\begin{aligned} v_{i,j}(\Delta t+1) = {} & wv_{i,j}(\Delta t) + c_1 r_1 (p_{i,j}(t) - \\ & x_{i,j}(\Delta t)) + c_2 r_2 (gbest_j(\Delta t) - x_{i,j}(\Delta t)) \end{aligned} \tag{7}$$

The position of particle i, x_i is updated as:

$$x_i(\Delta t+1) = x_i(\Delta t)+v_i(\Delta t+1) \quad (8)$$

The numbers of particles are initialized at random locations that correspond to feature subsets and then swarm towards promising areas via the global best solution so far and each particle's local best. The smallest subset with maximum quality is returned.

To begin with, the fuzzy rough quick reduct algorithm initializes the potential reduct to an empty set. By potential reduct, we mean the current best set of attributes. The first step in the fuzzy rough quick reduct algorithm is to define the equivalence classes. Let $A_1 = \{a_1\}$, $A_2 = \{a_2\}$, .., $A_n = \{a_n\}$ denotes the set of attributes and $Q = \{q\}$ denotes the decision class. The equivalence class is defined in Eq. 9 as:

$$U / A_1 = \{N_a, Z_a\}, ..., U / A_n = \{N_n, Z_n\} \quad (9)$$

And:

$$U / Q = \{\text{Tumor}; \text{Normal}\}$$

The next step is to compute the lower approximation of the sets A_1, A_2, ..., A_n. The lower approximation is computed using Eq. 9 as in Eq. 10:

$$\mu_{A1(\text{tumor})}(x) = \sup \min(\mu F(x), \inf \max\{1-\mu F(y), \mu\{\text{tumor}\}(y)\}) \quad (10)$$

The first fuzzy equivalence class of A_1, N_a is given in Eq. 11:

$$\min(\mu_{Na}(x), \inf \max\{1-\mu_{Na}(y), \mu_{(\text{tumor})}(y)\}) \quad (11)$$

After the computation of the lower approximation, the fuzzy positive region for each instance needs to be computed. This calculation is given as in Eq. 12 by:

$$\mu_{POS A(Q)}(x) = \sup \mu_{\underline{A}X}(x) \quad (12)$$

The final step is to calculate the degree of dependency. It is computed as in Eq. 13:

$$\gamma^1 A(Q) = \sum x \varepsilon U \, \mu_{POS A(Q)}(x) / |U| \quad (13)$$

The degree of dependency is computed for each of the attributes in the dataset. Only those dependency values that contribute to the increase in dependency degree are added to the potential reduct set. Whenever the addition of an attribute causes no increase in the dependency, at that point of time, the algorithm stops and prints the final minimal reduct set. Output: C is the reduced feature subset that would be obtained after applying the modified fuzzy rough quick reduct algorithm

RESULTS

This study is devoted to discuss the experimental setup of the proposed system. The raw data genes are subjected to normalization. Then, they are subjected to feature selection using the modified fuzzy rough quick reduct algorithm. The proposed method achieves dimensionality reduction using correlation coefficient on a particle swarm optimization search space. The number of features selected using the correlation based filter is tabulated in Table 3.

The subset of genes obtained after dimensionality reduction is subjected to Fuzzy Rough Quick Reduct algorithm. The number of features selected using the Fuzzy Rough Quick Reduct is tabulated in Table 4.

The performance of the proposed system is measured by running computer simulations on HP Workstation with intel xeon CPU with 3 GHz processor, 12 GB RAM on a Windows 7 operating system.

The Modified Fuzzy Rough Quick Reduct algorithm is implemented by using suitable functions as part of the open source statistical package R for all our datasets by setting the parameter values. The various steps performed to accomplish the task of feature selection using Modified Fuzzy Rough Quick Reduct are as follows:

- Step 1: Read the input file in comma separated format into R
- Step 2: Store the dataset as one single dataframe called decision. Table for further processing. This function takes 2 parameters namely dataset that contains instances and attributed as its rows and columns respectively and an integer value representing the index position of the decision attribute/class

Table 3: Number of genes selected using correlation based filter

Dataset name	No. of genes in the raw dataset	No. of genes obtained by correlation coefficient
ALL/AML	7,129	71
Lung Cancer	12,533	1,523
Ovarian Cancer	15,154	875

Table 4: Number of genes selected using modified fuzzy rough quick reduct

Dataset name	No. of genes obtained by correlation coefficient	No. of genes obtained from modified fuzzy quick reduct
ALL/AML	71	10
Lung Cancer	1,523	7
Ovarian Cancer	875	9

- Step 3: Compute the controls for the Fuzzy Rough Quick Reduct algorithm which is essential to compute the reduct in the next step

The controls are as follows. Aggregation: it is used to indicate the fuzzy indiscernibility relations. It is used for any fuzzy relations that determine the degree to which two objects are indiscernibility. Briefly, the indiscernibility relation is a relation that shows a degree of similarity among the objects.

For example, $R(x_i; x_j) = 0$ means the object x_i is completely different from x_j and otherwise, if $R(x_i; x_j) = 1$ while between those values we consider a degree of similarity. To calculate this relation, several methods have been implemented in this function which are approaches based on fuzzy tolerance, equivalence and t-equivalence relations.

- Tolerance: It indicates the Fuzzy Tolerance represented by suitable equations. For all our datasets, we use the one in Eq. 14:

$$R_a(x,y) = (1 - |a(x) - a(y)|) / (|a_{max} - a_{min}) \quad (14)$$

- The t-norm: Triangular norm that uses an operator called lukasiewicz (Implicator). The lukasiewicz is represented by max $(x_2+x_1-1; 0)$
- Type: We use the fuzzy quick reduct, the implementation in its original form
- Dependency: Determine the fuzzy lower and upper approximations using the implicator and t-norm function
- Compute the final reduct
- Store the result as a comma separated file for further classification

DISCUSSION

The performance of the proposed method was evaluated by using selected feature gene subsets from microarray cancer gene expression data using traditional classifiers. The entire dataset was used for the purpose of training and testing by using 10-fold cross validation strategy. Three binary cancer microarray gene expression datasets are used to compare and test the performance of feature selection using the proposed method and validated against correlation based feature selection and fuzzy rough quick reduct. The classifier accuracy of the reduced feature subset was analyzed using a 10-fold cross validation strategy and validated against traditional classifiers. In order to evaluate the performance of the classifier, the following parameters were used namely accuracy, precision, recall, F-measure and Region of Characteristic (ROC) Area. In order to compute the above parameters, it is essential to define certain terminologies namely:

- True positive (t_p)-equivalent with hit
- True negative (t_n)-Correct rejection
- False positive (f_p)-False alarm
- False negative (f_n)-Miss

Equation used to compute the accuracy of the classifier is given in Eq. 15:

$$\text{Accuracy} = (t_p + t_n) / (t_p + t_n + f_p + f_n) \quad (15)$$

The denominator value in Eq. 15 is called the total population size. Figure 3-5 depict the classifier accuracy

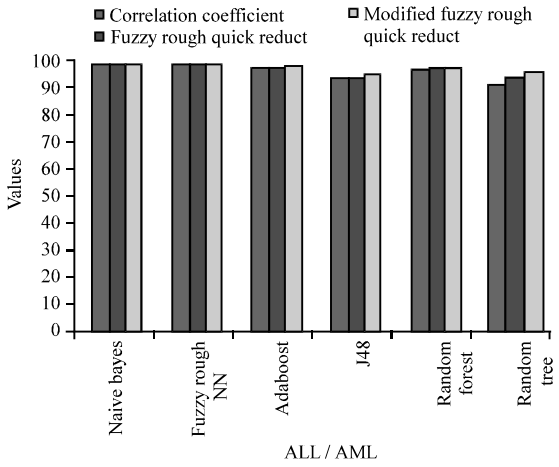

Fig. 3: Classifier accuracy for ALL/AML

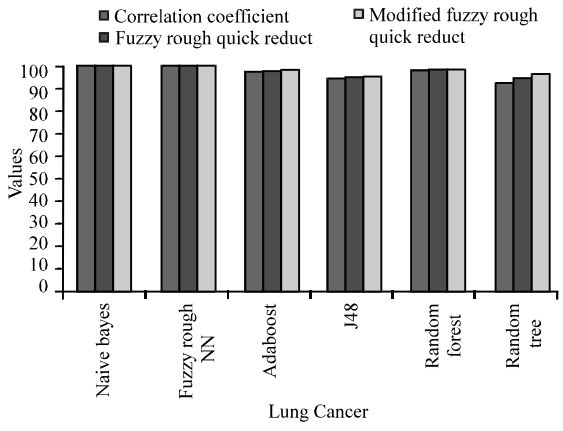

Fig. 4: Classifier accuracy for Lung Cancer

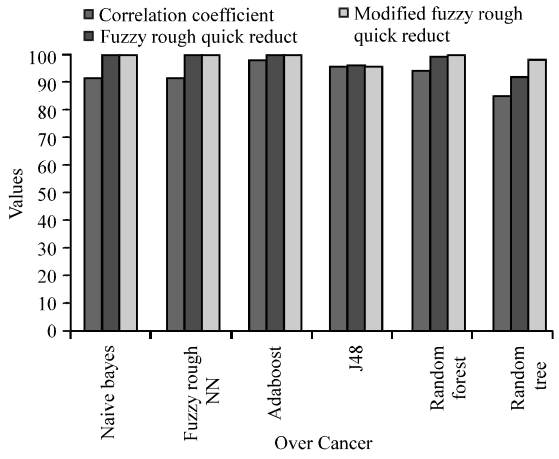

Fig. 5: Classifier accuracy for Ovarian Cancer

obtained by correlation coefficient, fuzzy rough quick reduct and modified fuzzy rough quick reduct algorithm. The comparative pictorial representation of the classifier accuracy of the proposed method with the traditional methods of feature selection like correlation coefficient and Fuzzy Rough Quick Reduct algorithm shows improved accuracy of the proposed method than other approaches.

Precision and recall are the two basic parameters used for evaluation in search strategies and based on understanding and measure of relevance. Precision also called the positive predictive value is the fraction of the retrieved instances that are relevant. Recall also called as sensitivity is the fraction of relevant instances that are retrieved. Equation used to compute the precision is given in Eq. 16:

$$\text{Precision} = t_p/(t_p + f_p) \quad (16)$$

Table 5-7 shows the comparative chart of the precision obtained by traditional classifiers using correlation coefficient, Fuzzy Rough Quick Reduct algorithm and our proposed approach for the 3 benchmarked datasets used for our study. All values marked in bold in the following tables signify the best level of precision, recall and F-measure that could be obtained by our proposed method compared to the feature selection approaches like correlation coefficient and fuzzy rough quick reduct. The formulae used to compute the recall also called sensitivity is given in Eq. 17:

$$\text{Recall} = t_p/(t_p + f_n) \quad (17)$$

Table 8-10 shows the comparative chart of the recall (Sensitivity) obtained by traditional classifiers using correlation coefficient, fuzzy rough quick reduct and our proposed approach for the 3 benchmarked datasets used for our study.

Table 5: Precision values for correlation coefficient, fuzzy rough quick reduct and modified fuzzy rough quick reduct ALL/AML dataset

ALL/AML (precision)				
Correlation coefficient (classifier)	Fuzzy rough quick reduct	Modified fuzzy rough quick	Reduct	Class
Naive bayes	0.978	0.978	0.939	ALL
	0.889	0.889	0.957	AML
	0.947	0.947	0.945	Weighted average
Fuzzy rough neural network				
	0.821	0.938	0.979	ALL
	0.938	0.917	0.960	AML
	0.862	0.930	0.972	Weighted average
Adaboost	0.936	0.936	0.918	ALL
	0.880	0.880	0.913	AML
	0.917	0.917	0.917	Weighted average
J48	0.929	0.915	0.936	ALL
	0.733	0.840	0.880	AML
	0.861	0.889	0.917	Weighted average
Random forest	0.938	0.938	0.939	ALL
	0.917	0.917	0.957	AML
	0.930	0.930	0.945	Weighted average
Random tree	0.857	0.929	0.894	ALL
	0.783	0.733	0.800	AML
	0.831	0.861	0.861	Weighted average

Table 6: Precision values for correlation coefficient, fuzzy rough quick reduct and modified fuzzy rough quick reduct Lung Cancer dataset

ALL/AML (precision)				
Correlation coefficient (classifier)	Fuzzy rough quick reduct	Modified fuzzy rough quick	Reduct	Class
Naive bayes	1.000	1.000	1.000	ADCA
	0.969	0.969	0.969	Mesothilioma
	0.995	0.995	0.995	Weighted average
Neural network	0.993	1.000	1.000	ADCA
	1.000	1.000	1.000	Mesothilioma
	0.995	1.000	1.000	Weighted average
Adaboost	0.980	0.980	0.987	ADCA
	0.966	0.966	0.967	Mesothilioma
	0.978	0.978	0.983	Weighted average
J48	0.967	0.967	0.980	ADCA
	0.839	0.839	0.848	Mesothilioma
	0.945	0.945	0.957	Weighted average
Random forest	0.974	0.939	0.980	ADCA
	1.000	0.957	1.000	Mesothilioma
	0.978	0.945	0.984	Weighted average
Random tree	0.966	0.967	0.974	ADCA
	0.743	0.839	0.931	Mesothilioma
	0.928	0.945	0.966	Weighted average

Normally precision and recall measures go hand-in-hand. The combined precison-recall measure is called the F-Measure. The F-measure is the harmonic mean of precision and sensitivity. In other words, F-score or F-measure in statistics is a measure of test's accuracy. It is computed as in Eq. 18:

$$\text{F-measure} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (18)$$

Table 7: Precision values for correlation coefficient, fuzzy rough quick reduct and modified fuzzy rough quick reduct Ovarian Cancer dataset

ALL/AML (precision)				
Correlation coefficient (classifier)	Fuzzy rough quick reduct	Modified fuzzy rough quick	Reduct	Class
Naive bayes	0.938	1.000	1.000	Cancer
	0.880	1.000	1.000	Normal
	0.917	1.000	1.000	Weighted average
Fuzzy rough neural network				
	0.902	1.000	1.000	Cancer
	0.937	1.000	1.000	Normal
	0.915	1.000	1.000	Weighted average
Adaboost	0.981	0.994	1.000	Cancer
	0.967	1.000	1.000	Normal
	0.976	0.996	1.000	Weighted average
J48	0.952	0.952	0.969	Cancer
	0.965	0.965	0.935	Normal
	0.957	0.957	0.957	Weighted average
Random forest	0.920	1.000	1.000	Cancer
	0.987	0.978	1.000	Normal
	0.944	0.992	1.000	Weighted average
Random tree	0.882	0.938	0.981	Cancer
	0.783	0.880	0.967	Normal
	0.846	0.917	0.976	Weighted average

Table 8: Recall values for correlation coefficient, fuzzy rough quick reduct and Modified Fuzzy rough quick reduct ALL/AML dataset

ALL/AML (precision)				
Correlation coefficient (classifier)	Fuzzy rough quick reduct	Modified fuzzy rough quick	Reduct	Class
Naive bayes	0.936	0.936	0.979	ALL
	0.960	0.960	0.880	AML
	0.944	0.944	0.944	Weighted average
Neural network	0.979	0.957	0.979	ALL
	0.600	0.880	0.960	AML
	0.847	0.931	0.972	Weighted average
Adaboost	0.936	0.936	0.957	ALL
	0.880	0.880	0.840	AML
	0.917	0.917	0.917	Weighted average
J48	0.830	0.915	0.936	ALL
	0.880	0.840	0.880	AML
	0.847	0.889	0.917	Weighted average
Random forest	0.957	0.957	0.979	ALL
	0.880	0.880	0.880	AML
	0.931	0.931	0.944	Weighted average
Random tree	0.894	0.830	0.894	ALL
	0.720	0.880	0.800	AML
	0.833	0.847	0.861	Weighted average

Table 10: Recall values for correlation coefficient, fuzzy rough quick reduct and modified fuzzy rough quick reduct Ovarian Cancer dataset

ALL/AML (precision)				
Correlation coefficient (classifier)	Fuzzy rough quick reduct	Modified fuzzy rough quick	Reduct	Class
Naive bayes	0.932	1.000	1.000	Cancer
	0.890	1.000	1.000	Normal
	0.917	1.000	1.000	Weighted average
	0.969	1.000	1.000	Cancer
	0.813	1.000	1.000	Normal fuzzy
	0.913	1.000	1.000	Weighted average
Adaboost	0.981	1.000	1.000	Cancer
	0.967	0.989	1.000	Normal
	0.976	0.996	1.000	Weighted average

Table 10: Continue

ALL/AML (precision)				
Correlation coefficient (classifier)	Fuzzy rough quick reduct	Modified fuzzy rough quick	Reduct	Class
J48	0.981	0.981	0.963	Cancer
	0.912	0.912	0.945	Normal
	0.957	0.957	0.957	Weighted average
Random forest	0.994	0.988	1.000	Cancer
	0.846	1.000	1.000	Normal
	0.941	0.992	1.000	Weighted average
Random tree	0.877	0.932	0.981	Cancer
	0.791	0.890	0.967	Normal
	0.846	0.917	0.976	Weighted average

Table 11: F-measure values for correlation coefficient, fuzzy rough quick reduct and Modified Fuzzy rough quick reduct ALL/AML dataset

ALL/AML (precision)				
Correlation coefficient (classifier)	Fuzzy rough quick reduct	Modified fuzzy rough quick	Reduct	Class
Naive bayes	0.957	0.957	0.958	ALL
	0.923	0.923	0.917	AML
	0.945	0.945	0.944	Weighted average
Fuzzy rough neural network				
	0.893	0.947	0.979	ALL
	0.732	0.898	0.960	A M L
	0.837	0.930	0.972	Weighted average
Adaboost	0.936	0.936	0.938	ALL
	0.88	0.880	0.875	AML
	0.917	0.917	0.916	Weighted average
J48	0.876	0.915	0.936	ALL
	0.8	0.840	0.880	AML
	0.85	0.889	0.917	Weighted average
Random forest	0.947	0.947	0.958	ALL
	0.898	0.898	0.917	A M L
	0.93	0.930	0.944	Weighted average
Random tree	0.857	0.876	0.894	ALL
	0.75	0.800	0.800	AML
	0.832	0.850	0.861	Weighted average

Table 11-13 shows the comparative chart of the F-measure obtained by traditional classifiers using correlation coefficient, fuzzy rough quick reduct and our proposed approach for the 3 benchmarked datasets used for our study.

The Receiver Operating Characteristic (ROC) curve can be plotted for each of the datasets considering the False Positive Rate (FPR) along the x-axis and True Positive Rate (TPR) along the y-axis of the graph. The ROC plots for the three datasets namely ALL/AML, Lung Cancer and Ovarian Cancer depicted in Fig. 6-8.

From figures and Tables it is clearly evident that the proposed method namely Modified Fuzzy Rough Quick Reduct algorithm selects much lesser number of genes (features) for all our three microarray cancer gene expression datasets, produces much higher accuracy has better values of precision, recall, f-measure and the ROC curves also shows that the classifier accuracy for the

Table 12: F-measure values for correlation coefficient, fuzzy rough quick reduct and modified fuzzy rough quick reduct Lung Cancer dataset

ALL/AML (precision)

Correlation coefficient (classifier)	Fuzzy rough quick reduct	Modified fuzzy rough quick	Reduct	Class
Naive bayes	0.997	0.997	0.997	ADCA
	0.984	0.984	0.984	Mesothilioma
	0.995	0.995	0.995	Weighted average
Fuzzy rough neural network				
	0.997	1.000	1.000	ADCA
	0.984	1.000	1.000	Mesothilioma
	0.994	1.000	1.000	Weighted average
Adaboost	0.987	0.987	0.990	ADCA
	0.933	0.933	0.951	Mesothilioma
	0.978	0.978	0.983	Weighted average
J48	0.967	0.967	0.973	ADCA
	0.839	0.839	0.875	Mesothilioma
	0.945	0.945	0.956	Weighted average
Random forest	0.987	0.958	0.990	ADCA
	0.931	0.917	0.949	Mesothilioma
	0.977	0.944	0.983	Weighted average
Random tree	0.953	0.967	0.980	ADCA
	0.788	0.839	0.900	Mesothilioma
	0.924	0.945	0.966	Weighted average

Table 13: F-measure values for correlation coefficient, fuzzy rough quick reduct and modified fuzzy rough quick reduct Ovarian Cancer dataset

ALL/AML (precision)

Correlation coefficient (classifier)	Fuzzy rough quick reduct	Modified fuzzy rough quick	Reduct	Class
Naive bayes	0.935	1.000	1.000	Cancer
	0.885	1.000	1.000	Normal
	0.917	1.000	1.000	Weighted average
	0.935	1.000	1.000	Cancer
Fuzzy rough neural network				
	0.871	1.000	1.000	Normal
	0.912	1.000	1.000	Weighted average
Adaboost	0.981	0.997	1.000	Cancer
	0.967	0.994	1.000	Normal
	0.976	0.996	1.000	Weighted average
J48	0.967	0.967	0.966	Cancer
	0.938	0.938	0.940	Normal
	0.956	0.956	0.957	Weighted average
Random forest	0.955	0.994	1.000	Cancer
	0.911	0.989	1.000	Normal
	0.940	0.992	1.000	Weighted average
Random tree	0.879	0.935	0.981	Cancer
	0.787	0.885	0.967	Normal
	0.846	0.917	0.976	Weighted average

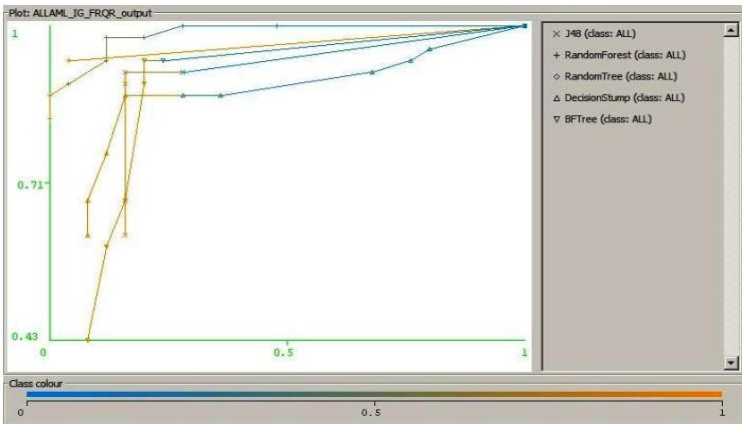

Fig. 6: The ROC plot for ALL/AML dataset for modified Fuzzy Rough Quick Reduct algorithm

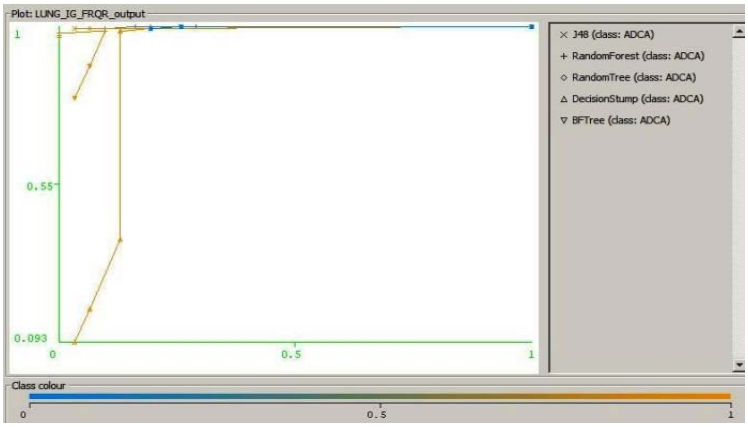

Fig. 7: The ROC plot for Lung Cancer dataset for modified Fuzzy Rough Quick Reduct algorithm

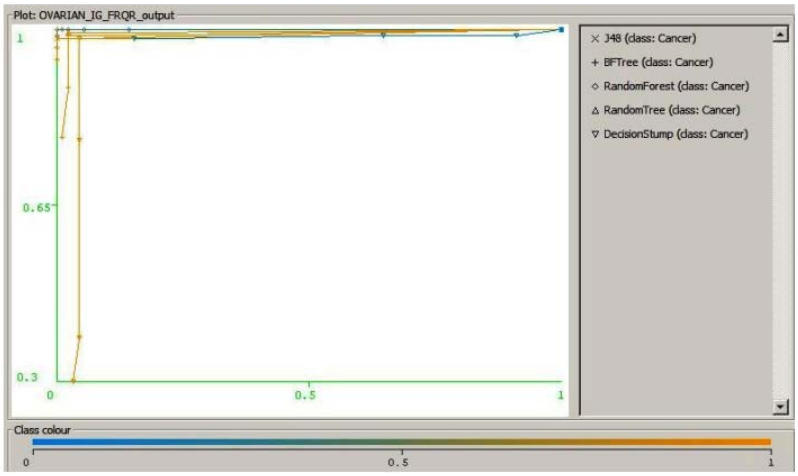

Fig. 8: The ROC plot for Ovarian Cancer dataset for modified Fuzzy Rough Quick Reduct algorithm

proposed method is much higher<other approaches like correlation coefficient and Fuzzy Rough Quick Reduct algorithm.

CONCLUSION

This study employed modified Fuzzy Rough Quick Reduct algorithm by using correlation based filter on a particle swarm optimization search space for dimensionality reduction followed by fuzzy rough quick reduct for feature selection. The performance of the different classifiers was analyzed using the reduced feature subset. The experimental results obtained by applying the modified fuzzy rough quick reduct method significantly reduced the total number of genes and increased the classifier accuracy compared to other traditional methods in literature. The proposed method could be applied to problems in other domain areas and could guide the research in the feature selection domain to varying dimensions in near future.

REFERENCES

Arunkumar, C. and S. Ramakrishnan, 2015. A comparative study of different classifiers on microarray cancer gene expression data. Aust. J. Basic Appl. Sci., 9: 145-151.

Bennet, J., C.A. Ganaprakasam and K. Arputharaj, 2014. A discrete wavelet based feature extraction and hybrid classification technique for microarray data analysis. Sci. World J., 2014: 1-9.

Bontempi, G., 2007. A blocking strategy to improve gene selection for classification of gene expression data. IEEE/ACM Trans. Comput. Biol. Bioinform., 4: 293-300.

Canedo, V.B., N.S. Marono, A.A. Betanzos, J.M. Benitez and F. Herrera, 2014. A review of microarray datasets and applied feature selection methods. Inf. Sci., 282: 111-135.

Chakraborty, D. and U. Maulik, 2014. Identifying cancer biomarkers from microarray data using feature selection and semisupervised learning. Transl. Eng. Health Med. IEEE. J., 2: 1-11.

Chen, D. and Y. Yang, 2014. Attribute reduction for heterogeneous data based on the combination of classical and fuzzy rough set models. Fuzzy Syst. IEEE. Trans., 22: 1325-1334.

Costa, P.D.J.F., H. Alonso and L. Roque, 2011. A weighted principal component analysis and its application to gene expression data. IEEE/ACM. Trans. Comput. Biol. Bioinf., 8: 246-252.

Dai, J.H. and Q. Xu, 2013. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. Applied Soft Comput., 13: 211-221.

Dash, S., B.N. Patra and B.K. Tripathy, 2012. Study of classification accuracy of microarray data for cancer classification using multivariate and hybrid feature selection method. IOSR. J. Eng., 2: 112-119.

Fu, L.M. and E.S. Youn, 2003. Improving reliability of gene selection from microarray functional genomics data. Inf. Technol. Biomed. IEEE. Trans., 7: 191-196.

Ghorai, S., A. Mukherjee, S. Sengupta and P.K. Dutta, 2011. Cancer classification from gene expression data by NPPC ensemble. IEEE/ACM. Trans. Comput. Biol. Bioinf., 8: 659-671.

Hall, M.A., 1999. Correlation-Based Feature Selection for Machine Learning. University of Waikato Press, New Zealand, Pages: 178.

Hu, Q., L. Zhang, S. An, D. Zhang and D. Yu, 2012. On robust fuzzy rough set models. Fuzzy Syst. IEEE. Trans., 20: 636-651.

Hu, X. and N. Cercone, 1995. Learning in relational databases: A rough set approach. Comput. Intell., 11: 323-338.

Kar, S., K.D. Sharma and M. Maitra, 2015. Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique. Expert Syst. Appl., 42: 612-627.

Kumar, C.A. and S. Ramakrishnan, 2014. Binary classification of cancer microarray gene expression data using extreme learning machines. Proceedings of the 2014 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), December 18-20, 2014, IEEE, Coimbatore, India, pp: 1-4.

Lazar, C., J. Taminau, S. Meganck, D. Steenhoff and A. Coletta *et al*., 2012. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM. Trans. Comput. Biol., 9: 1106-1119.

Leung, Y. and Y. Hung, 2010. A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification. Comput. Biol. Bioinf. IEEE/ACM. Trans., 7: 108-117.

Liu, Y., 2008. Detect key gene information in classification of microarray data. EURASIP. J. Adv. Signal Proc., 1: 1-10.

Maulik, U. and D. Chakraborty, 2014. Fuzzy preference based feature selection and semisupervised SVM for cancer classification. NanoBiosci. IEEE. Trans., 13: 152-160.

Skowron, A., 1995. Extracting laws from decision tables: a rough set approach. Comput. Intell., 11: 371-388.

Yang, C.S., L.Y. Chuang, C.H. Ke and C.H. Yang, 2008. A hybrid feature selection method for microarray classification. IAENG. Int. J. Comput. Sci., 35: 285-290.

Ziarko, W., 1993. Variable precision rough set model. J. Comput. Sys. Sci., 46: 39-59.