Future Computing and Informatics Journal

Volume 3
Issue 1 (2018) *Volume 3, Issue 1, 2018*

Article 10

2018

Attribute selection using fuzzy roughset based customized similarity measure for lung can cer microarra y gene expression data

C. Arunkumar
Dept. of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India, c_arunkumar@cb.amrita.edu

S. Ramakrishnan
Dept. of Information Technology, Dr. Mahalingam College of Engineering and Technology, Pollachi, India, ram_f77@yahoo.com

Available online at www.sciencedirect.com

ScienceDirect

Future Computing and Informatics Journal 3 (2018) 131−142
http://www.journals.elsevier.com/future-computing-and-informatics-journal/

Attribute selection using fuzzy roughset based customized similarity measure for lung cancer microarray gene expression data

C. Arunkumar [a,*], S. Ramakrishnan [b]

[a] Dept. of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India
[b] Dept. of Information Technology, Dr. Mahalingam College of Engineering and Technology, Pollachi, India

Received 30 June 2017; revised 21 November 2017; accepted 8 February 2018
Available online 20 February 2018

Abstract

Microarray gene expression data plays a prominent role in feature selection that helps in diagnosis and treatment of a wide variety of diseases. Microarray gene expression data contains redundant feature genes of high dimensionality and smaller training and testing samples. This paper proposes a customized similarity measure using fuzzy rough quick reduct algorithm for attribute selection. Information Gain based entropy is used to reduce the dimensionality in the first stage and the proposed fuzzy rough quick reduct method that defines a customized similarity measure for selecting the minimum number of informative genes and removing the redundant genes is employed at the second stage. The proposed method is evaluated using leukemia, lung and ovarian cancer gene expression datasets on a random forest classifier. The proposed method produces 97.22%, 99.45% and 99.6% classifier accuracy on leukemia, lung and ovarian cancer gene expression datasets respectively. The research study is carried out using the R open source software package. The proposed method shows substantial improvement in the performance with respect to various statistical parameters like classification accuracy, precision, recall, f-measure and region of characteristic compared to available methods in literature.

Keywords: Feature selection; Information gain; Fuzzy rough quick reduct; Random forest

1. Introduction

Cancer is considered to be a deadly disease across the globe. Traditional methods of diagnosis are time consuming and error prone as they depend entirely on human judgement. So machine learning methods and algorithms aid in the early diagnosis and treatment of the disease thereby increasing the survival rate in the area of biomedical and bioinformatics [8,24]. Feature selection has gained importance in the recent years. Microarray databases have grown in rows and columns in the recent years [22]. The irrelevant features present in the high dimensional dataset occupies large amount of memory space that deteriorates the performance of the learning algorithm [46]. Higher dimensionality of the microarray dataset motivates researchers to perform feature selection using a variety of approaches. Microarray gene expression data finds its application in the diagnosis and treatment of different types of cancer. Some challenges faced are large number of feature genes, fewer numbers of samples and lack of proper validation since gene expression data is prone to outliers and noise [5]. Microarray technology is widely adopted for measurement and monitoring of the gene expression activation levels that find its application in the diagnosis and treatment of a wide variety 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

* Corresponding author.
E-mail addresses: c_arunkumar@cb.amrita.edu (C. Arunkumar), ram_f77@yahoo.com (S. Ramakrishnan).
Peer review under responsibility of Faculty of Computers and Information Technology, Future University in Egypt.

https://doi.org/10.1016/j.fcij.2018.02.002

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 kentridge biomedical repository, NCBI, Genbank, Array Express, Gene Expression Omnibus, and Stanford Microarray. The microarray dataset that is used in our study is of the format as in Table 1, where FG_1, FG_2, FG_3, …,FG_N indicates the Gene ID, 1,2, …,N indicates the instances which represents the data of each sample and the N^{th} column indicates the class and the numerical values represent the gene expression levels. In this case, all values in the sample table lie between −1 and 1 which means that the data is normalized.

Uncertainty, inaccuracy and fuzziness of data can be dealt using a valuable tool called the rough set theory. It reduces the number of features without any additional information by determining the dependencies of the data [18,27,49]. A minimal attribute subset is derived from a superset and is called the reduct. This reduct needs to exhibit the same discernibility as the superset preserving the semantics so as to minimize loss of information during feature selection. The highlight of this approach is that it selects the most significant genes/features from the superset without transformation of data [25,37]. Several methods have been proposed to determine the reducts. Fuzzy rough sets encapsulates the distinct and related concepts of fuzziness and indiscernibility. Knowledge uncertainty is considered to be one of the reasons for this occurrence. Fuzzy rough attribute reduction is achieved by computing the dependency relation. Some of the works in Refs. [4,13,18,44]; focus majorly on the computation of attribute reducts by different approaches. On the compact computational domain, the computational efficiency of the algorithm is improved [13]. The concept of information entropy is used to compute the fuzzy rough reduct in Ref. [33]. An improved fuzzy discernibility matrix is developed in Ref. [18]. In Ref. [44], a theoretical foundation was laid for attribute reduction in Ref. [19]. A study was undertaken through approach of discernibility matrix [31].

The traditional fuzzy rough quick reduct algorithm computes the reducts by employing three different similarity measures as proposed by Refs. [42,20,34]. Though they produce reducts comparatively smaller in size than the proposed similarity measure, they have two key disadvantages. The similarity measures are complex in nature. The computation of reducts results in elimination of relevant genes in the final reducts that result in reduced classifier accuracy. For these two reasons, a customized similarity measure is proposed using Shannon entropy based information gain filter for dimensionality reduction and feature selection is performed using customized fuzzy rough quick reduct (FRQR) algorithm.

Table 1
Format of gene expression data.

Instance	FG_1	FG_2	FG_3	FG_N	Class
1	−0.286	−0.095	0.213	−0.802	normal
2	0.659	−0.672	−0.023	−0.861	tumour
3	−0.800	0.089	0.134	1.000	tumour
N	−0.973	−0.786	1.000	−0.913	normal

The key objective of this research paper is to develop a minimal reduct for leukemia, lung and ovarian cancer microarray gene expression datasets using a customized similarity measure for fuzzy rough quick reduct approach that computes all the reducts. The proposed measure is compared with fuzzy similarity measure proposed in Refs. [42,20,34]; of attribute reduction. Our algorithm proves to be effective after analysing various statistical parameters like classification accuracy, precision, recall, F-measure and Region of Characteristic (ROC). This paper is organized as follows. The related work and background study carried out by researchers in the field of rough set and fuzzy rough set is presented in Section 2. Section 3 presents the customized similarity measure in fuzzy rough quick algorithm for attribute selection. The experimental results and discussion is outlined in Section 4 and the conclusion is presented in Section 5.

2. Related work and background of the research work

Heuristic based techniques are used to implement several feature selection methods in rough set theory. 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 in Ref. [40]. Attribute reduction using positive region is discussed in Ref. [10]. The same kind of approach with target decision unchanged is discussed in Ref. [14]. The concept of using information entropy to search 'reduct' in a rough set model is discussed in Ref. [41]. The above concept is expanded to approximate reduct that could be used for a number of feature reduction methods is discussed in Ref. [50]. The concept of feature selection using fuzzy rough set has been discussed by several authors that could be summarized in Table 2 below.

2.1. Background of the research work

2.1.1. Rough set approach

The concepts of incompleteness and uncertainty that arises in numerous domains of research can be dealt using a new mathematical tool called the Rough Set (RS) theory invented in 1982 by Pawlak that uses the concepts of set models, approximation space, lower and upper approximation. Preliminary or additional information is not required by the rough set and this serves as a key advantage during the implementation phase. Roughset theory finds its applications in computation of the minimal reduct by comparing equivalence or similarity measures in attribute reduction. The computation of the minimal reduct set depends on the degree of dependency measure and should be noted that the reduced subset produces the same degree of dependency as the unreduced set [11].

2.1.2. Information gain in roughset theory

Given a decision system $DS = \langle \cup, C_a \cup D_a, V, f \rangle$, where C_a and D_a represents the set of conditional and decision attributes respectively, V is the union of attribute domains, $V = \cup_{a \in A} V_a$

Table 2
Summary of feature selection approaches using Fuzzy Rough Set.

Reference	Approach	Classifier	Dataset	Remarks
[33]	Using SAT to compute rough and fuzzy rough reducts	JRip	Lung, Heart	Performs better than Rough Set Attribute reduction
[49]	Computation of reducts using fish swarm algorithm	Decision rules	Lung	Quick coverage, strong search capability, finds minimal reducts efficiently with competitive performance
[3]	Customized fuzzy rough quick reduct approach for feature selection	Naive Bayes, Fuzzy rough neural network, Adaboost, J48, Random forest, Random Tree	Leukemia, Lung cancer, ovarian cancer	Reduction in number of gene subsets, improved accuracy
[35]	Neighbourhood approximation and grouping	JRip,IBK	Leukemia, colon, Lymphoma	Reduction in execution time, improvement in performance in terms of subset size
[43]	Genetic algorithm based computerized recognition of autism gene expression data	SVM	Breast cancer	Identifies features best associated with the disease, good performance
[29]	Attribute selection using maximum relevance and significance criteria	K-NN, SVM, C.4.5	Breast cancer, leukemia, colon, lung cancer	Discriminative genes are selected from high dimensional datasets
[30]	Rough set feature gene selection for microarray data based on maximum relevance and significance criterion	K-NN, SVM	Breast Cancer, leukemia, Lung Cancer, and Leukemia	Improvement in gene selection, feasibility and effectiveness of the proposed method
[32]	Fuzzy Lower Approximation	Specialized classifiers	Wine dataset	Increased robustness, effective k-mean and k-median
[6]	Attribute reduction for heterogenous data	SVM	UCI dataset	Inconsistency and attribute reduction
[21]	Fuzzy gain ratio based attribute selection	SVM, C4.5	SRBCT, colon, hepatocellular carcinoma	Effective attribute selection approach
[11]	Rough set and PSO based hybrid feature selection	Naive Bayes, BayesNet, k-star	Breast Cancer	Increase in predictive accuracy
[47]	Information entropy based feature selection using novel fuzzy roughset in mixed data	LSVM,KSVM,CART	Breast Cancer	Tradeoffs between Better feature selection and predictive accuracy
[48]	Rough set based technique for learning fuzzy rules from fuzzy samples	ID3	UCI dataset	Reduced Precision of the learning algorithm
[26]	Attribute reduction in inconsistent decision tables	C4.5, RBF-SVM	Breast Cancer, Heart	Improved robustness, efficiency and feasibility
[15]	Feature selection and classification on medical database based on threshold fuzzy entropy	RBF	Breast Cancer	Increase in classification accuracy

where V_a is called the value set of attribute a, called the domain of a. $f : \cup \times A \rightarrow V$ is a decision function. $B \subseteq C_a$, $\cup|B = \{x_1, \ldots x_n\}$ *and* $\cup|D_a = \{y_1, \ldots y_m\}$, the conditional entropy of D_a conditioned to B is defined in Eq. (1) as

$$H(D_a|B) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\left|X_i \cap Y_j\right|/|\cup| \times \log\left|X_i \cap Y_j\right|/X_i \quad (1)$$

The mutual information of B and D_a is defined as $I(B;D_a) = H(D_a) - H(D_a|B)$ [21].

2.1.3. Tolerance rough set

The uncertain information present in the boundary regions are determined by using a distance metric in tolerance rough set [11]. These tolerance rough sets are defined using similarity measures of the lower and upper approximations of feature values [39]. The relaxation of transitivity constraint introduces further degree of indiscernibility among equivalence classes. Equivalence class grouping of objects in a traditional roughset is done if the values of the attributes are equal. In case of continuous data, the values differ because of noise and hence this requirement might seem to be too strict for continuous data [20] [34]; [23]. Let $\cup$ be a non-empty universe of discourse and $F(\cup \times \cup)$ be the fuzzy power set on $(\cup \times \cup)$. R is called a fuzzy relation on $(\cup \times \cup)$ if $R \in F(\cup \times \cup)$, where $R(x,y)$ measures the strength of relationship between $x \in \cup$ and $y \in \cup$. Let R be a fuzzy relation on $(\cup \times \cup)$. R is reflexive if $R(x,x) = 1$ for any $x \in \cup$; R is symmetric if $R(x,y) = R(y,x)$ for any $x, y \in \cup$ and R is T-transitive if $R(x,y) \geq T(R(x,z), R(z,y))$ for a triangular norm T and any $x, y, z \in \cup$. Furthermore, R is called a T-similarity relation if R is reflexive, symmetric and T-transitive. Specially, if T = min; R is called a fuzzy equivalence relation [47].

2.1.4. Similarity measure

The tolerance rough set approach defines a similarity measure for each attribute whose standard measure is represented using three relations as Eqs. (2)–(4) as in Refs. [42,20,34].

$$\mu_{R_a}(x,y) = 1 - a(x) - a(y)/a_{\max} - a_{\min} \quad (2)$$

$$\mu_{R_a}(x,y) = \exp^{\left(-\left((a(x)-a(y))^2/2\sigma_a^2\right)\right)} \quad (3)$$

$$\mu_{R_a}(x,y) = \max(\min((a(y) - a(x) + \sigma_a/\sigma_a), ((a(x) - a(y) + \sigma_a/\sigma_a), 0))) \quad (4)$$

where 'a' is the attribute under consideration, a_{max} and a_{min} denote the maximum and minimum values for the features taken, σ represents the standard deviation for the attribute and $\mu_{R_a}(x,y)$ denotes the degree of similarity that exists between objects 'x' and 'y' for feature 'a'. The above three equations (2)–(4) satisfy the two properties of fuzzy sets namely Reflexivity in Eq. (5) and Symmetricity in Eq. (6) are represented as

$$R(x,x) = 1 \tag{5}$$

$$R(x,y) = R(y,x) \tag{6}$$

2.1.5. Reduct and fuzzy rough quick reduct

The two main goals of attribute reduction are removal of redundant feature genes from the raw dataset and preserving the quality of the reduced feature subset. The information system needs to be maintained in a concise form in a majority of applications. The original raw dataset should be represented minimally using the concept of a reduct represented by a minimal subset R for the initial feature set C such that for a given set of features D, $\gamma_R(D) = \gamma_C(D)$ where γ represents the dependency degree. From the literature, R is a minimal subset if $\gamma_{R-\{a\}}(D) \neq \gamma_R(D)$ for all $a \in R$. Removal of features from the subset R would not be possible without affecting the dependency degree. A given dataset might have many reduct sets, and the collection of all reducts is denoted in Eq (7) by

$$R_{all} = \{X | X \subseteq C, \gamma_X(D) = \gamma_C(D); \gamma_{X-\{a\}}(D) \neq \gamma_X(D)\} \tag{7}$$

The core reduct is obtained by taking the intersection of all the sets in R_{all} and elimination of features becomes difficult without introducing more contradictions to the representation of the dataset. An ideal solution is to determine a single element of the reduct which represents a reduct of minimal cardinality as given in Eq. (8) as

$$R_{\min} \subseteq R_{all}: R_{\min} = \{X | X \in R_{all}, \forall Y \in R_{all}, |X| \leq |Y|\} \tag{8}$$

The fuzzy rough quick reduct algorithm computes the minimal reduct set from the several subsets available. The fuzzy rough quick reduct algorithm computes the fuzzy indiscernibility (identification of similar attributes in the raw dataset) which eliminates the superfluous attributes. Let $DS = (\cup, A)$ be a decision system. A decision system DS is represented as $DS: T = (\cup, A \cup \{d\})$ where'd' represents the decision attribute and the elements of A constitute the conditional attributes. The concept of indiscernibility is central to roughset theory. For any $B \subseteq A$, there is an associated indiscernibility relation $IND(B)$ represented in Eq. (9) as

$$IND(B) = [(x,y) | \forall a \in B, f(a,x) = f(a,y)] \tag{9}$$

The rough membership function μ is denoted as μ_X^B as represented in Eqs (10) and (11)

$$\mu_X^B : U \rightarrow [0,1] \tag{10}$$

$$\mu_X^B = \frac{|[x]_B \cap X|}{|[x]_B|} \tag{11}$$

where $P(x \in X|u)$ and 'u' is the equivalence class of the indiscernibility relation.

Then the fuzzy lower and upper approximations are computed using the formulae considering the value of $\pi = 1$ and μ represents the membership function. Eq. (12) represents the fuzzy lower approximation and Eq. (13) represents the fuzzy upper approximations.

$$\underline{R}X = \cup\{Y \in U/R : Y \subseteq X\} \tag{12}$$

$$\overline{R}X = \cup\{Y \in U/R : Y \cap X \neq \phi\} \tag{13}$$

The ordered pair $\langle \underline{R}X, \overline{R}X \rangle$ is called the roughset of X with respect to the equivalence relation $IND(B)$. Eqs (12) and (13) can also be rewritten as represented below in Eqs. (14) and (15) as

$$\underline{R}X = \left\{x \in \cup \middle| [x]_B \subseteq X\right\} \tag{14}$$

$$\underline{R}X = \left\{x \in \cup \middle| [x]_B \cap X \neq \phi\right\} \tag{15}$$

The lower and upper approximation of a set X with respect to $IND(B)$ is the set of all objects which certainly belongs to X and possibly belongs to X respectively with respect to $IND(B)$. The similarity measures as represented in Eqs. (2)–(4) and the positive region for each attribute are computed. The degree of dependency of the set of attributes is computed using the lukasiewicz triangular norm which serves a basis for computing the minimal reduct set using the fuzzy rough quick reduct (FRQR) algorithm.

3. Proposed approach to feature selection using customized similarity measure for fuzzy rough quick reduct algorithm

The key objective of this research work is to identify the prominent and informative genes that causes cancer by eliminating the redundant genes and to determine the best genes/features subset for identifying the different types of cancer. This objective is met by ranking the genes using information gain filter and removing the redundant genes by using a customized similarity measure for fuzzy rough quick reduct algorithm. Then, the reduced feature subset is analysed for accuracy using random forest classifier. High dimensional data suffers from the problem of "curse of dimensionality". Hence dimensionality reduction is one of the prerequisites for most of the high dimensional data analysis that must be carried out beforehand to make machine learning process more effective. Some of the advantages of applying the dimensionality reduction technique using information gain filter is that it selects a subset of informative feature genes from the original dataset without altering its properties, reduces the execution time of the fuzzy rough quick reduct algorithm and it can efficiently distinguish between different decision classes. The proposed framework for the proposed customized similarity measure in fuzzy rough quick reduct algorithm is shown in Fig. 1.

In our earlier work [3], we have used existing correlation based filter and modified fuzzy rough quick reduct algorithm

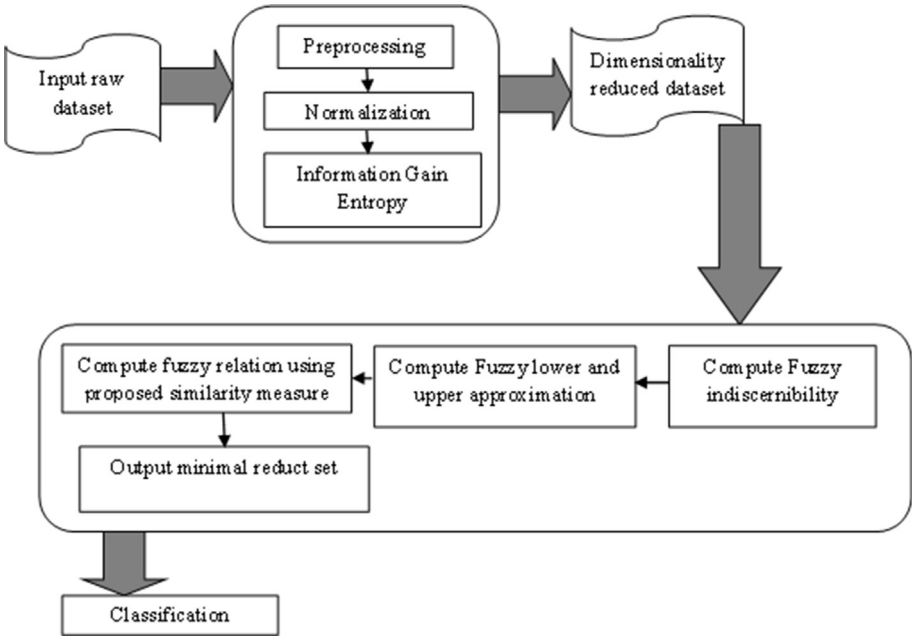

Fig. 1. Proposed Framework for the proposed customized similarity measure in fuzzy rough quick reduct algorithm.

using Particle Swarm Optimization (PSO) to compute the minimal reduct set. Though our previous work provided reasonable classification accuracy, it requires more number of genes. Also, PSO used for dimensionality reduction consumed extra computational time. Hence there is scope for reducing the number of feature genes and improving the classification accuracy with reduced computational time. In order to meet out the above said objectives, the proposed algorithm using Shannon entropy based information gain filter for dimensionality reduction and proposed customized similarity measure in fuzzy rough quick reduct is devised for computing the minimal reduct set. In order to analyse the performance of this proposed work, experiments across multiple datasets are conducted in a stringent setup using 10-fold cross validation and an in-depth performance analysis is performed for the proposed method. It can be found from Table 9 that the proposed method produces better classification accuracy with lesser number of feature genes. Since the proposed method uses information gain filter for dimensionality reduction and proposed customized similarity measure in fuzzy rough quick reduct algorithm for computing the minimal reduct set, it is computationally less intensive compared to the earlier work that used PSO based technique which is computationally more intensive.

Table 3
Summary of gene expression datasets used.

Dataset	Number of genes	Class	References
Leukemia gene expression data	7129	ALL/AML	[17,26,28,12,45,5,36]
Lung cancer	12533	ADCA/Mesothelioma	[17,28,5]
Ovarian cancer	15154	Tumour/Normal	[17,28,5]

The following section presents the most relevant aspects of the proposed customized fuzzy rough quick reduct algorithm.

3.1. Feature ranking using information gain entropy

The process of feature ranking is performed using Shannon entropy. Information Gain filter is applied on the raw dataset by adopting the algorithm given below. The Information Gain is determined using the following algorithm: Given a decision system $DS = \langle \cup, C_a \cup D_a, V, f \rangle$, where $\cup$ represents the non-empty set of finite attributes often called as universe of discourse, C_a and D_a represents the set of conditional and decision attributes respectively, V is the union of attribute domains, $V = \cup_{a \in A} V_a$ where V_a is called the value set of attribute a, called the domain of a. $f : \cup \times A \rightarrow V$ is a decision function. $B \subseteq C_a$, $\forall x \in C_a - B$, the gain of attribute 'x', Gain (x, B, d) can be defined as in Eq (16) below

$$\begin{aligned} Gain(x,B,D_a) &= I(B \cup \{x\}; D_a) - I(B; D_a) \\ Gain(x,B,D_a) &= H(D_a|B) - H(D_a|B \cup \{x\}) \end{aligned} \tag{16}$$

If $B = \phi, Gain(x,B,D_a) = H(D_a) - H(D_a|\{x\}) = I(\{x\}; D_a)$, An attribute 'x' in an attribute set B for a decision attribute D_a gains significant importance whenever the value of Gain (x, B, D_a) is higher. Mutual information gain based algorithm for attribute selection can be described as follows. Information Gain is computed using the following steps listed below:

Input: Let R_D represent the raw dataset obtained after the process of normalization. R_D is subjected to dimensionality reduction using Information Gain filter. Let C_a and D_a set of conditional and decision attributes produced by R_D.

Step 1. Let $B = \phi$, where B represents the selected attributes from the raw dataset

Table 4
Comparison of accuracy and number of genes – Raw dataset Vs Dimensionality reduced dataset.

Dataset	Number of genes in the raw dataset	Number of genes obtained by shannon entropy based information gain (Dimensionality reduced dataset)	FRQR on raw dataset		FRQR on dimensionality reduced dataset	
			Number of genes obtained	Classifier accuracy (%)	Number of genes obtained	Classifier accuracy (%)
Leukemia gene expression data	7129	1025	15	90.28	7	97.22
Lung cancer	12533	4981	14	95.58	6	99.45
Ovarian cancer	15154	6237	12	97.23	9	99.60

Step 2. The significance of the condition attribute 'x', Gain (x, B, Da)is computed for every attribute, $x \in C_a - B$
Step 3. The attribute that maximizes the gain value Gain(x, B, Da is noted as 'x'.
Step 4. If Gain(x, B, Da) > 0, then $B \leftarrow B \cup \{x\}$ goto Step 2, else goto Step 5;
Step 5. The set B is the selected attributes that possess entropy value > 0 [21].

Output: The reduced feature subset that is ordered based on the ranking of features. All features with an entropy value of zero are eliminated.

3.2. Obtaining quick reduct using fuzzy tolerance and triangular norm

The reduced feature subset obtained after applying the Information Gain filter is subjected to a customized fuzzy rough quick reduct method. This method removes the redundant genes and produces a reduct that contains most prominent genes that cause different types of cancer. Let B represent the Information Gain, R_D denotes the raw dataset and γ^1 denotes the dependency function. This function is used to decide attributes to add or ignore to the current reduct set. The termination condition for the algorithm is the constant value of dependency obtained even after addition of any remaining attribute to the current reduct set. Let C_a and D_a denote the set of conditional and decision attributes produced by R_D. C denotes the subset otherwise called the reduct set that would be generated by this algorithm. The algorithm for this approach is given as below.

$B \leftarrow R_D$

$C \leftarrow \{\}; \gamma^1_{best} = 0; \gamma^1_{prev} = 0$

while $\gamma^1_{best} \neq \gamma^1_{prev}$
$T \leftarrow C$
$\gamma^1_{prev} = \gamma^1_{best}$
foreach $x \in (C_a - C)$
if $\gamma^1_{C \cup \{x\}}(B) > \gamma^1_T(B)$
$T \leftarrow C \cup \{x\}$
$\gamma^1_{best} = \gamma^1_T(B)$
$C \leftarrow T$
return C

The different steps involved in the computation of the reduct set using a customized similarity measure for the fuzzy rough quick reduct algorithm [16] is as follows:

i) Load the dataset from CSV file
ii) List all instances
iii) Calculate equivalence classes of decision attribute
iv) Calculate tolerance of each attribute
v) Calculate the indescernible set
vi) Create list of relation matrix for each attribute
vii) Calculate positive region of each attribute
viii) Calculate dependency degree

3.2.1. Compute fuzzy indiscernibility

Fuzzy indiscernibility is the key concept in computing the minimal reduct set using FRQR algorithm. The degree of similarity between features could be determined using the concept of indiscernibility relation. For instance, the degree of similarity lies between 0 and 1. If $R(x_a, x_b) = 0$, then the two features are dis-similar, if $R(x_a, x_b) = 1$, the two features are similar and intermediate values exhibit some degree of similarity. The computation of the fuzzy indiscernibility is done using a variety of methods namely fuzzy tolerance, equivalence and T-equivalence relations. This paper proposes the

Table 5
Statistical parameters for leukemia, lung and ovarian cancer gene expression datasets.

Dataset	Similarity measure	No of attributes selected	Classifier accuracy (%)	Precision	Recall	F-measure	TP	FN	FP	TN
Leukemia	Similarity measure proposed in Ref. [42]	8	94.44	0.949	0.944	0.943	21	0	4	47
	Proposed similarity measure	7	97.22	0.972	0.972	0.972	23	1	2	46
Lung cancer	Similarity measure proposed in Ref. [42]	7	97.24	0.972	0.972	0.972	149	1	4	27
	Proposed similarity measure	6	99.45	0.995	0.994	0.995	149	0	1	31
Ovarian cancer	Similarity measure proposed in Ref. [42]	9	97.23	0.973	0.972	0.972	161	1	6	85
	Proposed similarity measure	9	99.60	0.996	0.996	0.996	161	0	1	91

Table 6
FPR, TPR values for leukemia gene expression data, lung and ovarian cancer gene expression datasets.

Dataset	Similarity measure proposed in Ref. [42]			Customized similarity measure (proposed method)		
	TPR	FPR	AUC	TPR	FPR	AUC
Leukemia gene expression data	0.944	0.067	0.919	0.972	0.041	0.985
Lung cancer	0.972	0.108	0.988	0.994	0.001	0.999
Ovarian cancer	0.972	0.106	0.990	0.996	0.002	1.000

Table 7
Comparison of similar work carried out by other researchers.

Materials and Methods	Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy [47]	A threshold fuzzy entropy based feature selection for medical database classification [15]	Our proposed work
Dataset used with number of features/genes in parenthesis	SPECT (22), Lymphography (18),Promotergenesequences (57), Zoo (16), Wine (13), Libras movement (90), Wisconsin prognostic breast cancer (33), Wisconsin diagnostic breast cancer (30), Horse colic (22), Statlog (13), Credit approval (15), German creditdata (20)	Medical dataset - Wisconsin Breast Cancer (9), Pima Indians Diabetes (8), Heart-Statlog (13), Hepatitis (19) and Cleveland Heart Disease (13)	Microarray gene expression dataset − Leukemia gene expression data (7129), lung cancer (12534), ovarian cancer (15154)
Primary focus of research in the paper	Fuzzy roughset based information entropy	Fuzzy entropy	FRQR with customized similarity measure with Shannon entropy based information gain for dimensionality reduction
Average number of feature genes in the final reduct set (in parenthesis)	SPECT (1), Lymphography (5),Promotergenesequences (1), Zoo (4), Wine (6), Libras movement (15), Wisconsin prognostic breast cancer (1), Wisconsin diagnostic breast cancer (3), Horse colic (4), Statlog (7), Credit approval (2), German creditdata (7)	Medical dataset - Wisconsin Breast Cancer (6), Pima Indians Diabetes (4), Heart-Statlog (7), Hepatitis (12) and Cleveland Heart Disease (8)	Microarray gene expression dataset − Leukemia gene expression data (7), lung cancer (6), ovarian cancer (9)
Classifiers used and classifier accuracy reported	LSVM (80.16), KSVM (83.67), CART (79.83)	RBF network (86%)	Random forest (98.76)

computation of the indiscernibility relation by using a customized similarity measure as listed in (22).

3.2.2. Using proposed similarity measure

Let the tabular representation of an information system be denoted as $C_a = (\cup, C_a)$. $\cup$, called as the universe of discourse denotes finite number of objects that are part of a non-empty set and C_a denotes a finite set of attributes that belong to a non-empty set such that $C_a : \cup \rightarrow V_x$ for every $x \in C_a$. The set V_x is the set of values that attribute 'x' may take. An information system transforms itself into a decision system if it contains a decision attribute for each object. It is represented as $C_a = (\cup, C_a \cup \{D_a\})$, where $D_a \notin C_a$ represents the decision attribute. The conditional attributes are obtained

Table 8
Comparison of proposed method with UCI datasets.

Dataset name	Number of features in raw dataset	[16]		Proposed method	
		Number of features selected	Classifier accuracy	Number of features selected	Classifier accuracy
Wine	13	6	95.88	5	97.3
SPECTF	44	8	75.97	2	76.56
SONAR	60	12	74.47	5	76.38
PIMA Indian Diabetes	8	7	75.46	4	78.26

Table 9
Feature selection approaches – Number of genes and classifier accuracy – A comparison (Feature subset size is indicated in parenthesis).

Feature Selection methods	Leukemia gene expression data	Lung cancer	Ovarian cancer
CFS-LFS [1]	95.83 (52)	98.81 (163)	99.47 (36)
CFS [2]	93.10 (113)	96.94 (274)	97.62 (641)
PLSDR [26]	97.1 (20)	–	–
BCGS [28]	94.1 (35)	91.2 (34)	98.8 (26)
GEM [12]	91.5 (3)	–	–
IWSS [45]	94.4 (8)	–	–
BDE-SVM$_{rankf}$ [17]	82.40 (6)	98.00 (3)	95.00 (3)
IWSS-MB-NB [45]	97.1 (7)	–	–
DRF0 [5]	91.18 (13)	98.66 (17)	100 (16)
IRLDA [36]	97 (72)	–	–
CFS-PSO-FRQR [3]	92.59 (10)	98.07 (7)	98.88 (9)
BDE-SVM$_{rank}$ [17]	82.4 (7)	98 (3)	100 (3)
Similarity measure proposed in Ref. [42]	94.44 (8)	97.24 (7)	97.23 (9)
Customized similarity measure (Proposed method)	**97.22 (7)**	**99.45 (6)**	**99.60 (9)**

Bold signifies that our proposed method produces higher classification accuracy compared to available methods in literature.

as part of C_a. The subset of features P induces a fuzzy similarity relation represented as R_P in Eq. (17):

$$\mu_{R_P}(x,y) = \cap_{x\in P}\{\mu_{R_a}(x,y)\} \tag{17}$$

The equivalence classes for fuzzy similarity measure and triangular norm are tuned. When more than one feature is taken into consideration, the defined similarities must be combined so that an overall similarity exhibited between different genes could be measured. For a subset of features, P, it could be achieved in two different ways using (18) and (19) [20,34]; namely:

$$(x,y)\in SIM_{P,\tau} iff \prod_{a\in P} SIM_a(x,y) \geq \tau \tag{18}$$

$$(x,y)\in SIM_{P,\tau} iff \sum_{a\in P} SIM_a(x,y)/|P| \geq \tau \tag{19}$$

where τ is called the global similarity threshold that determines the required level of similarity for inclusion within the tolerance class. The tolerance classes that are generated by a given similarity relation for an object x is defined in (20) as

$$SIM_{P,\tau}(x) = \{y\in\cup | (x,y)\in SIM_{P,\tau}\} \tag{20}$$

$\mu_{R_P}(x,y)$ for objects (x,y) denotes the degree of similarity between x and y using the attribute values of R_P. The classical indiscernibility relation for a qualitative attribute is defined in (21) as

$$\mu_{R_P}(x,y) = \{1 \text{ if } R_P(x) = R_P(x) \text{ and } 0 \text{ if } R_P(x) \neq R_P(y)\} \tag{21}$$

The fuzzy similarity measure for all our datasets is represented by using (22) as

$$\mu_{R_a}(x,y) = \left(1 - abs\left(\left(a\left(x^2\right) + a\left(y^2\right) - 2\times a(x)\times a(y)\right)\right)\right) \tag{22}$$

The proposed similarity measure in Eq. (22) satisfies the two properties of fuzzy sets namely Reflexivity (Eq. (5)) and Symmetricity (Eq. (6)).

3.2.3. Fuzzy lower approximation based feature selection

The lower approximation is generalized by using a triangular norm (t-norm) and an implicator. The lukasiewicz fuzzy Implicator is represented by Eq. (23) and the Lukasiewicz t-norm by Eq (24) as

$$\min(1 - x_1 + x_2, 1) \tag{23}$$

$$\max(x + y - 1, 0) \tag{24}$$

The fuzzy B-lower approximation of the fuzzy set A in ∪ is given as in Eq. (25) as

$$(R_B \downarrow A)(y) = \inf_{x\in\cup} \tau_{imp}(R_B(x,y), A(x)) \tag{25}$$

$R_B \downarrow A$ denotes the set of elements necessarily belonging to a particular set and is said to possess strong membership.

3.2.4. Computation of positive region

Let P and Q be equivalence relations over ∪, then the positive region is defined by (26) as

$$\mu_{POS_{R_P}(Q)}(x) = \sup_{X\in\cup/Q}\left(\mu_{\underline{R_P X}}(x)\right) \tag{26}$$

The fuzzy similarity relation is denoted by R_P that is induced by a subset of features denoted by P. The information of features in P is used to find if the attributes fall into the positive region provided they have sufficient information to showcase indiscernibility.

3.2.5. Computation of dependency degree

Determining the dependency among features is a key task in computing the minimal reduct set using FRQR. The functional dependency relationship is said to exist between two attributes P and Q if P depends totally on Q.

For $P, Q \subseteq C_a$ Q depends on P in a degree $k (0 \leq k \leq 1)$ denoted by $P \Rightarrow_k Q$, if

$$k = \lambda_p(Q) = \sum_{x\in\cup} \mu_{POS_{R_P}(Q)}(x)/|\cup| \tag{27}$$

The dependency that exists between the conditional and decision attributes is denoted by $\lambda_p(Q)$ and is termed as the quality of approximation. The value of 'k' determines the dependency value and it lies between 0 and 1. A value of 0 indicates no dependency and 1 indicates total dependency. Any value between 0 and 1 indicates partial dependency.

3.2.6. Computation of the minimal reduct set and classification

The change in the dependency value determines the significance of an attribute. The change in dependency value is computed whenever an attribute is added to the reduct set. If the change is more, higher the significance of the feature else it would be discarded from the reduced subset. Let the reduced

feature subset that would be obtained after applying the proposed customized similarity measure to attribute selection be represented by C. Using C, random forest classifier is applied and analysed to predict the classification accuracy, precision, recall, F-measure and region of characteristic. A usual and adequate measure in microarray data is the accuracy of the classifier. This might be affected because of the fact that the microarray data contains very few numbers of training and testing samples. This problem could be solved by using the 10-fold cross validation strategy that splits the training data into 10-subsets of the same size. The classification accuracy is computed by averaging the estimations obtained from each of the 10 different subsets [38]; [7].

4. Experimental results and discussion

The feature selection is carried out using R open source software package and the classification is done using the random forest classifier. The various statistical parameters like classification accuracy, precision, recall, F-measure and region of characteristic for the proposed method is analysed and evaluated.

4.1. Dataset description

The various statistical parameters like classification accuracy, precision, recall and region of characteristic for the proposed method is analysed and evaluated using leukemia, lung and ovarian cancer gene expression datasets downloaded from the kentridge biomedical repository. Table 3 presents the summary of gene expression datasets used for this study.

Acute Lymphocytic Leukemia (ALL) is a type of cancer that originates from the immature lymphocytic cells. Acute Myeloid Leukemia (AML) starts its development in the bone marrow other than the lymphocytic cells and moves quickly into the blood. There are 58 instances of ALL and 14 instances of AML. The binary dataset consists of 7129 genes taken from 72 samples. Lung cancer also called the lung carcinoma is a malignant tumour caused by uncontrolled cell growth in the lung tissue. The two classes of lung cancer namely ADCA and mesothelioma samples numbering 245 are collected. Mesothelioma samples contain more than 50% tumour cells and ADCA consists of both metastatic and primary malignancies were taken from the colon and breast. Tumour blocks were used to obtain total RNA with the help of suitable reagents. The hybridization of cRNA was performed using probe arrays. Since few samples (64 in number) revealed artefacts, they were discarded and 181 samples were used for further analysis [9]. The lung cancer gene expression dataset consists of 12 533 feature genes taken from 181 patient samples (150 samples of adenocarcinoma and 31 samples of malignant pleural mesothelioma). Ovarian cancer forms in the ovary. These abnormal cells have higher chances of spreading to other parts of the body. The ovarian cancer gene expression dataset consists of 15 154 feature genes taken from 253 samples. There are 162 instances of cancerous samples and 91 instances of normal samples. All the three binary datasets consists of raw data.

4.2. Performance analysis

Initially, the raw dataset is subjected to normalization on a scale of [-1 1]. 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 [-1, 1] and dimensionality reduction and feature selection are performed on the normalized dataset. The fuzzy rough quick reduct (FRQR) algorithm is applied on the raw dataset and dimensionality reduced dataset. The results are tabulated in Table 4. It can be inferred that the number of feature genes selected and the classifier accuracy are lesser compared to the feature selection methods applied on the dimensionality reduced datasets. Dimensionality reduction serves two purposes namely removal of redundant genes and improvement in classification accuracy. Hence the raw dataset is subjected to dimensionality reduction using the Information Gain method. Table 4 below shows the number of feature genes obtained after the process of dimensionality reduction using Shannon entropy based information gain filter.

Suitable experiments have been performed on three binary cancer microarray gene expression datasets namely leukemia, lung and ovarian cancer gene expression datasets. The classification accuracy reported for leukemia, lung and ovarian cancer gene expression datasets using random forest classifier when the raw datasets are used is 86.11%, 81.94% and 92.89% respectively. Similarly, the classification accuracy is computed for the dimensionality reduced datasets. The predictive accuracy is estimated to be 88.89%, 95.03% and 94.86% for leukemia, lung and ovarian cancer gene expression datasets respectively. The experiments are carried out using the similarity measures proposed by Refs. [42,20,34]; and our proposed customized similarity measure.

The proposed feature selection approach that uses a customized similarity measure is classified using 10-fold cross validation strategy on a random forest classifier. The results obtained for the leukemia, lung and ovarian cancer gene expression datasets are tabulated in Table 5. The various statistical parameters are analysed and tabulated. They include precision, recall, F-measure, true positive (TP), true negative (TN), false positive (FP) and false negative (FN). The relevance measure could be well understood using the three basic parameters namely precision, recall and F-measure. They are widely used for evaluation in search strategies. Precision is the fraction of the retrieved instances that are relevant and recall or sensitivity is the fraction of relevant instances that are retrieved [3]. The F-measure is widely used in statistics to obtain an accurate measurement of a test's accuracy. Precision, recall and F-measure are computed using the formulae given in (28), (29) and (30) respectively.

$$\text{Precision} = TP/(TP + FP) \tag{28}$$

$$\text{Recall} = TP/(TP + FN) \tag{29}$$

$$F-measure = 2 \times (precision \times recall)/(precision + recall) \quad (30)$$

4.3. Comparison based on ROC curve

Additionally the Region of Characteristic (ROC) is drawn for all the three datasets under study. Additional conceptual information can be visualized for the classifier accuracy metric using the ROC plot. The formula to compute the FPR and TPR is given below in (31) and (32):

$$FPR \approx FP/(TN + FP) \quad (31)$$

$$TPR \approx TP/(TP + FN) \quad (32)$$

The FPR and TPR values are tabulated in Table 6 as under for the leukemia, lung and ovarian cancer gene expression datasets where FPR indicates False Positive Rate and TPR indicates True Positive Rate. The area under the curve (AUC) is based on the concept of probability distribution that produces a single value from the ROC curve. It represents the probability that a randomly chosen positive sample would be ranked higher by a classifier than a randomly chosen negative sample [17].

The x-axis in the plots below represents the FPR and the y-axis represents the TPR. ROC curves are plotted for the leukemia gene expression data and lung cancer datasets in Figs. 2–5. Similar results are obtained for the ovarian cancer dataset.

4.4. Comparison with state-of-the art feature selection approaches

The following key distinctions are noted when our proposed work is compared with "Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy" and "A threshold fuzzy entropy based feature selection for medical database classification" and the comparison is tabulated in Table 7.

The key inferences to be noted are as follows:

- The referred papers focus on the medical/other datasets that are low dimensional in nature whereas our proposed research work focuses on high dimensional cancer microarray gene expression data.

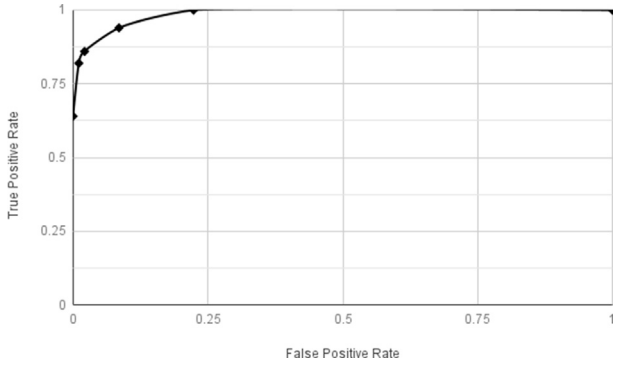

Fig. 2. ROC plot for leukemia gene expression data dataset – Similarity measure proposed in Ref. [42].

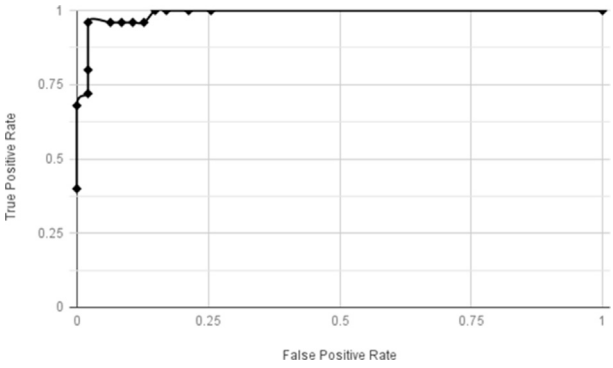

Fig. 3. ROC plot for leukemia gene expression data dataset – proposed customized similarity measure.

- Since our datasets are high dimensional in nature, dimensionality reduction using Shannon entropy based information gain filter is applied on the normalized dataset before computing the final reduct set in order to reduce the computational time.
- Our proposed method of customized similarity measure has reduced the number of feature genes and has also contributed to increase in classification accuracy using random forest classifier with 10-fold cross validation strategy compared to the reduction of features in the referenced papers.

Our proposed approach selects 7 feature genes (0.1% of the total genes) from the leukemia gene expression data dataset, 6 feature genes (0.05% of the total genes) from the lung cancer gene expression dataset and 9 feature genes (0.06% of the total genes) from the ovarian cancer gene expression dataset. Our proposed method is implemented on the different datasets used in the paper [16]. The comparative results are presented in Table 8 below and our proposed method performs better for other datasets as well.

Our earlier research works are compared with the proposed feature selection approach. The performance of the proposed method using a customized similarity measure is compared with similar approaches namely correlation based filter and extreme learning machines classifier [2], correlation based filter on a linear forward selection search strategy [1]

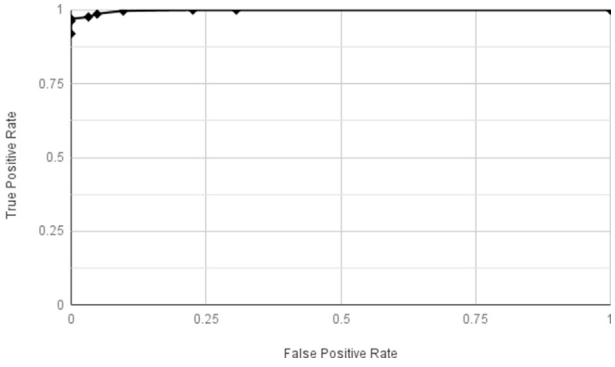

Fig. 4. ROC plot for lung cancer dataset – Similarity measure proposed in Ref. [42].

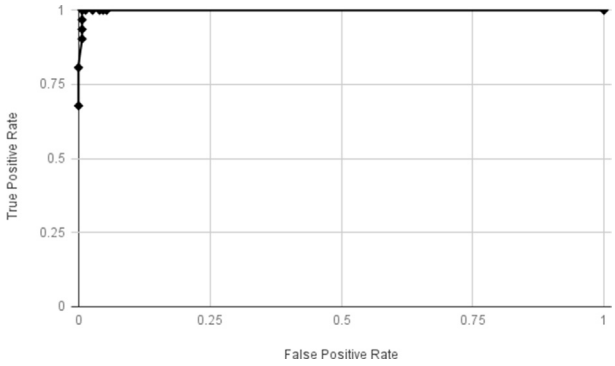

Fig. 5. ROC plot for lung cancer dataset − proposed customized similarity measure.

and a hybrid approach to FRQR algorithm that combines correlation based filter on a particle swarm optimization strategy and uses FRQR algorithm for feature selection [3]. With reference to our previous works carried out in feature selection and classification, the proposed method shows substantial improvement in terms of gene reduction and classification accuracy. The comparison of the classification accuracies on different feature selection techniques is tabulated in Table 9.

5. Conclusion

Fuzzy roughset concept is used extensively in attribute reduction and selection of microarray gene expression data because of its high dimensionality. This paper proposed a customized similarity measure for attribute selection using fuzzy rough quick reduct algorithm in leukemia, lung and ovarian cancer microarray gene expression datasets. Dimensionality reduction is performed using Shannon entropy based information gain filter. The proposed approach is evaluated on different datasets using the existing similarity measures on a random forest classifier. Experimental results based on the proposed customized similarity measure exhibits higher classification accuracy and shows promising results compared to the ones available in literature. The same could be extended for diagnosis and prevention of other diseases as well in future.

Conflict of interest statement

On behalf of all the authors, I, the corresponding author (Arunkumar Chinnaswamy) state that there is no conflict of interest.

Acknowledgement

We would like to express our deep sense of gratitude to the Management team and Chairperson of the Department of Computer Science and Engineering, Amrita University, India and also the Management, Secretary and Principal of Dr.Mahalingam College of Engineering and Technology, Pollachi, India for supporting us to carry out this research work.

References

[1] Arunkumar C, Ramakrishnan SA. Comparative study of different classifiers on microarray gene expression data. Aust J Basic Appl Sci 2015; 27:145−51.
[2] Arunkumar C, Ramakrishnan S. Binary classification of cancer microarray gene expression data using extreme learning machines. In: Proceedings of 2014 IEEE International conference on computational intelligence and computing research; 2014. p. 1−4.
[3] Arunkumar C, Ramakrishnan S. Modified Fuzzy rough quick reduct algorithm for feature selection in cancer microarray data. Asian J Inf Technol 2016;15:199−210.
[4] Bhatt RB, Gopal M. On fuzzy rough sets approach to feature selection. Pattern Recogn Lett 2005;26:965−75.
[5] Bolon Canedo V, Sanchez Marono N, Alonso-Betanzos A. Distributed feature selection, an application to microarray data classification. Appl Soft Comput 2015;30:136−50.
[6] Chen Degang, Yang Yanyan. Attribute reduction for heterogeneous data based on the combination of classical and fuzzy rough set models. IEEE Trans Fuzzy Syst 2014;22:1325−34.
[7] Duval B, Hao JK. Advances in metaheuristics for gene selection and classification of microarray data. Briefings Bioinf 2009;1:127−41.
[8] Ahmad Fadzil, Mat Isa Nor Ashidi, Hussain Zakaria, Khusairi Osman Muhammad, Sulaiman Siti Noraini. A GA-based feature selection and parameter optimization of an ANN in diagnosing breast cancer. Pattern Anal Appl 2015;18:861−70.
[9] Gavin Gordon J, Roderick Jensen V, Hsiao Li-Li, Steven Gullans R, Joshua Blumenstock E, Ramaswamy Sridhar, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Research 2002;62(17):4963−7.
[10] Grzymala Busse J. An algorithm for computing a single covering. Managing uncertainty in expert systems. Netherlands: Kluwer Academic Publishers; 1991. p. 66.
[11] Hannah Inbarani H, Ahmad Taher Azar Jothi G. Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. Comput Meth Progr Biomed 2014;113:175−85.
[12] Hernandez JCH, Duval B, Hao JK. A genetic embedded approach for gene selection and classification of microarray data. Evolutionary computation, machine learning and data mining in bioinformatics. Springer; 2007. p. 90−101.
[13] Hu QH, Xie ZX, Yu DR. Hybrid attribute reduction based on a novel fuzzy rough model and information granulation. Pattern Recogn 2007;40: 3509−21.
[14] Hu XH, Cercone N. Learning in relational databases, a rough set approach. Comput Intell 1995;11:323−38.
[15] Jaganathan P, Kuppuchamy R. A threshold fuzzy entropy based feature selection for medical database classification. Comput Biol Med 2013;43: 2222−9.
[16] Anaraki Javad Rahimipour, Eftekhari Mahdi. Improving fuzzy-rough quick reduct for feature selection. In: Proceedings of 19th Iranian conference on electrical engineering; 2011. p. 1−6.
[17] Apolloni Javier, Leguizamon Guillermo, Alba Enrique. Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. Appl Soft Comput 2016;38:922−32.
[18] Jensen R, Shen Q. New approaches to fuzzy-rough feature selection. IEEE Trans Fuzzy Syst 2009;17:824−38.
[19] Jensen R, Shen Q. Fuzzy-rough attributes reduction with application to web categorization. Fuzzy Set Syst 2004;141:469−85.
[20] Jensen R, Shen Q. Rough set based feature selection, a review in rough computing. Inf Sci 2007:70−107.
[21] Dai Jianhua, Xu Qing. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumour classification. Appl Soft Comput 2013;13:211−21.
[22] Liang Jiye, Wang Feng, Dang Chuangyin, Qian Yuhua. An efficient rough feature selection algorithm with a multi-granulation view. Int J Approx Reason 2012;53:912−26.

[23] Jothi G, Hannah Inbarani H. Hybrid Tolerance Rough Set–Firefly based supervised feature selection for MRI brain tumor image classification. Appl Soft Comput 2016;46:639–51.
[24] Kourou Konstantina, Exarchos Themis P, Exarchos Konstantinos P, Karamouzis Michalis V, Fotiadis Dimitrios I. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17.
[25] Yong Liu, Wenliang Huang, Yunliang Jiang, Zhiyong Zeng. Quick attribute reduct algorithm for neighbourhood rough set model. Inf Sci 2014;271:65–81.
[26] Li Min, Shang Changxing, Feng Shengzhong, Fan Jianping. Quick attribute reduction in inconsistent decision tables. Inf Sci 2014;254:155–80.
[27] Summair Raza Muhammad, Qamar Usman. An incremental dependency calculation technique for feature selection using rough sets. Inf Sci 2016; 343:41–65.
[28] Pang S, Havukkala I, Hu Y, Kasabov N. Classification consistency analysis for bootstrapping gene selection. Neural Comput Appl 2007; 16(6):527–39.
[29] Maji Pradipta, Garai Partha. On fuzzy-rough attribute selection, criteria of max-dependency, max-relevance, Min-Redundancy, and max-significance. Appl Soft Comput 2013;13:3968–80.
[30] Maji Pradipta, Paul Sushmita. Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data. Int J Approx Reason 2011;52:408–26.
[31] He Qiang, Wu Congxin, Chen Degang, Zhao Suyun. Fuzzy rough set based attribute reduction for information systems with fuzzy decisions. Knowl Base Syst 2011;24:689–96.
[32] Hu Qinghua, Zhang Lei, An Shuang, Zhang David, Yu Daren. On robust fuzzy rough set models. IEEE Trans Fuzzy Syst 2012;20:636–51.
[33] Jensen Richard, Tuson Andrew, Shen Qiang. Finding rough and fuzzy-rough set reducts with SAT. Inf Sci 2014;255:100–20.
[34] Jensen Richard, Shen Qiang. Tolerance-based and fuzzy-rough feature selection. In: Proceedings of the IEEE International fuzzy systems conference; 2007. p. 1–6.
[35] Jensen Richard, Parthalain Neil Mac. Towards scalable fuzzy–rough feature selection. Inf Sci 2015;323:1–15.
[36] Sharma A, Paliwal KK, Imoto S, Miyano S. A feature selection method using improved regularized linear discriminant analysis. Mach Vis Appl 2014;25(3):775–86.
[37] Teng Shu-Hua, Lu Min, Feng Yang A, Zhang Jun, Nian Yongjian, He Mi. Efficient attribute reduction from the viewpoint of discernibility. Inf Sci 2016;326:297–314.
[38] Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95:14–8.
[39] Skowron A, Stepaniuk J. Tolerance approximation spaces. Fundam Inf 1996;27:245–53.
[40] Skowron A. Extracting laws from decision tables, a rough set approach. Comput Intell 1995;11:371–88.
[41] Slezak D. Foundations of entropy-based bayesian networks, theoretical results & rough set based extraction from data. In: Proceedings of the 8th international conference on information processing and management of uncertainty in knowledge-based systems; 2000. p. 248–55.
[42] Stepaniuk J. Optimizations of rough set model. Fundam Inf 1998; 36(2–3):265–83.
[43] Latkowski Tomasz, Osowski Stanislaw. Computerized system for recognition of autism on the basis of gene expression microarray data. Comput Biol Med 2015;56:82–8.
[44] Tsang CCE, Chen DG, Yueng SD, Lee WTJ, Wang XZ. Attribute reduction using fuzzy rough sets. IEEE Trans Fuzzy Syst 2008;16:1130–41.
[45] Wang A, Chen NG, Yang J, Li L, Alterovitz G. Incremental wrapper based gene selection with markov blanket. In: Proceedings of IEEE international conference on bioinformatics and biomedicine (BIBM), IEEE; 2014. p. 74–9.
[46] Shu Wenhao, Shen Hong. Incremental feature selection based on rough set in dynamic incomplete data. Pattern Recogn 2014;47:3890–906.
[47] Zhang Xiao, Mei Changlin, Chen Degang, Li Jinhai. Feature selection in mixed data, A method using a novel fuzzy rough set-based information entropy. Pattern Recogn 2016;56:1–15.
[48] Wang Xizhao, Eric Tsang CC, Zhao Suyun, Chen Degang, Daniel Yeung S. Learning fuzzy rules from fuzzy samples based on rough set technique. Inf Sci 2007;177:4493–514.
[49] Chen Yumin, Zhu Qingxin, Xu Huarong. Finding rough set reducts with fish swarm algorithm. Knowl Base Syst 2015;81:22–9.
[50] Ziarko W. Variable precision rough set model. J Comput Syst Sci 1993; 46:39–59.

Mr.C.Arunkumar, received the B.E. degree in Computer Science and Engineering in 2004 from the Bharathiar University, Coimbatore, and the M.Tech. degree in Computer Science and Engineering in 2006 from Vellore Institute of Technology university, Vellore. He has 12 years of teaching experience at undergraduate and post graduate level. He has published 20 papers in international journals and conferences. He is the reviewer for Elsevier - Knowledge Based Systems, Computer Methods and Programs in Biomedicine and Springer – International Journal of Fuzzy Systems.

Dr.S.Ramakrishnan, received the B.E. (ECE) in 1998, a M.E. (CS) in 2000 and PhD degree in Information and Communication Engineering from Anna University, Chennai in 2007. He is a Professor and the Head of IT Department, Dr. Mahalingam College of Engineering and Technology, Pollachi. He has 17 years of teaching experience and 1 year industry experience. He has published 152 papers and 8 books.
Dr.S.Ramakrishnan is an Associate Editor for IEEE Access and he is a Reviewer of 25 International Journals including 7 IEEE Transactions, 5 Elsevier Science Journals, 3 IET Journals, ACM Computing Reviews, Springer Journals, Wiley Journals, etc. He is in the editorial board of 7 International Journals. He is a Guest Editor of special issues in 3 International Journals including Telecommunication Systems Journal of Springer. His areas of research include digital image processing, information security, and soft computing.