Iranian Journal of Fuzzy Systems

Volume 16, Number 2, (2019), pp. 165-182

University of Sistan and Baluchestan

A hybrid filter-based feature selection method via hesitant fuzzy and rough sets concepts

M. Mohtashami[1] and M. Eftekhari[2]

[1,2] *Department of computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran*

m.mohtashami88@gmail.com, m.eftekhari@uk.ac.ir

Abstract

High dimensional microarray datasets are difficult to classify since they have many features with small number of instances and imbalanced distribution of classes. This paper proposes a filter-based feature selection method to improve the classification performance of microarray datasets by selecting the significant features. Combining the concepts of rough sets, weighted rough set, fuzzy rough set and hesitant fuzzy sets for developing an effective algorithm is the main contribution of this paper. The mentioned method has two steps, in the first step, four discretization approaches are applied to discretize continuous datasets. Also, a primary subset of features is selected by combining of weighted rough set dependency degree and information gain via hesitant fuzzy aggregation approach. In the second step, a significance measure of features (defined by fuzzy rough concepts) is employed to remove redundant features from primary set. The Wilcoxon Signed Ranked test (A Non-parametric statistical test) is conducted for comparing the presented method with ten feature selection methods across seven datasets. The results of experiments show that the proposed method is able to select a significant subset of features and it is an effective method in the literature in terms of classification performance and simplicity.

Keywords: Rough set, Weighted Rough set, Information gain, Discretization, Hesitant fuzzy set.

1 Introduction

Feature selection is one of the most important fields in machine learning, data mining, and pattern recognition. Feature selection is utilized as a preprocessing step to deal with high dimensional datasets since there is a lot of unnecessary information in this type of datasets that makes classification step computational. The most feature selection approaches attempt to remove irrelevant and redundant features. Feature selection methods can be categorized in filter-based, wrapper-based and embedded-based approaches [3, 1]. The filter based feature selection methods utilize a metric to select features and eventually one subset with highest metric value is selected. In the following, some filter methods are described briefly.

- **Correlation-based feature selection (CFS)** [14] is a method which ranks the features with correlation-based criteria. CFS starts with an empty set and selects the features that are highly correlated with class labels (relevancy) and have low correlation with other selected features (redundancy).
- **ReliefF** [17] is an extended form of Relief algorithm. The original Relief selects a sample randomly and finds its nearest neighbors from different classes. The values of features in selected sample are compared to nearest neighbors and the relevancy of each feature is updated. The ReliefF focuses on features that have different values in various classes and same values in same classes.
- The **min-Redundancy and Max-Relevance (mRMR)** [29] selects the features that are maximally dissimilar to each other and have the highest relevancy to the class attribute. The mRMR methods are based on mutual information in both relevancy and redundancy criteria.

Corresponding Author: M. Eftekhari
Received: July 2017; Revised: March 2018; Accepted: October 2018.

- **Rough set (RS)** [16] is a tool to compute the dependency between features (conditional attributes) and class labels (decision attribute). The dependency degree can be used as a measure of relevance to rank features that are more relevant (dependent) with targets.

Furthermore, a combination of filter based feature selection algorithms with each other or with other feature selection methods is used to provide hybrid feature selection approaches. Hybrid feature selection methods are an aggregation of multi feature selection algorithms by an appropriate aggregation approach. Some interesting hybrid methods can be found in references [26, 18, 33]. In last decade, there have been some interesting extensions on Rough set theory (RST) [8] for attribute reduction (feature selection). RS is a robust tool to deal with uncertain, vague and inexact problems. RST introduced by Z. Pawlak in 1982 [28], then this theory has been considered by many researchers in recent years. Classical rough set [16] just handles datasets with nominal attribute values. In fact common real world datasets have real-value (continuous) attributes, there are two ways to take action on real-value datasets. In the first way, continuous dataset should be discretized, this way is very sensitive to discretization method and it is inevitable that a lot of information will be deleted in discretization process. In the second way, The RS should be changed to deal with real-valued data and there are some developed methods based on rough set to handle this type of data such as fuzzy rough set, rough set neighborhood and etc. Fuzzy rough set [15, 22, 36] and rough set neighborhood [35, 5, 6] utilize fuzzy relations and distance measures to compute equivalence classes.

Recently, researchers have utilized filter based approaches for feature selection in high dimensional microarray datasets. In the following, the latest filter-based feature selection algorithms are mentioned. Ebrahimpour and Eftekhari [10] proposed a feature selection algorithm inspired from correlation-based feature selection (CFS), named MRMR-HFS. This algorithm selects features by the ensemble of ranking methods and similarity measures. The aggregation concepts of hesitant fuzzy sets (HFS) are utilized to aggregate the ranking methods and similarity measures. Therefore, Ebrahimpour et al. [11] propounded an algorithm to release high dimensional datasets from computational search methods, called Reduced Row Echelon Form Feature Selection (RREF-FS). This algorithm uses two steps for feature selection. In the first step, features are sorted through their importance in the dataset via information gain and in the second step, linear independent features from the dataset are selected by Reduced Row Echelon Form (RREF) concept. In [23] the authors presented fuzzy equivalence partition matrix (FEPM) to approximate the true marginal and joint distributions of continuous gene expression values. The fuzzy equivalence partition matrix is based on the theory of fuzzy rough sets, where each row of the matrix represents a fuzzy equivalence partition that can automatically be derived from the given expression values. The authors in [24] established a feature selection method based on rough set theory. This method selects a set of features from microarray data by maximizing the relevance and significance of the selected features (RSMRMS). P. Maji [21] introduced a feature selection algorithm based on a rough hypercuboid approach that called RHEPM-MRMDMS. This method selects a set of features from a data set by maximizing the relevance, dependency, and significance of the selected features. By introducing the concept of the hypercuboid equivalence partition matrix, a novel representation of degree of dependency of sample categories on features is proposed to measure the relevance, dependency, and significance of features in approximation spaces. The equivalence partition matrix also offers an efficient way to calculate many quantitative measures to describe the inexactness of approximate classification.

This paper is structured as follows. Basic concepts and notations on rough set theory and its developed methods are expressed in Section 2. In Section 3, a feature selection method based on weighted rough set and basic rough set with multi discretization approaches is proposed. In Section 4, experimental results are given. Finally, this paper is concluded in Section 5.

2 Preliminaries

In this section, basic concepts and notations on rough set theory and its developed methods such as weighted rough set and some notations on feature selection will be reviewed.

2.1 Rough set basic concepts

The rough set structure is defined by an information system (information table). Let $IS = \langle U, A, V, f \rangle$ be an information system and $U = \{x_1, x_2, \ldots, x_n\}$ be a set of instances (universe of discourse) in which x_i is the ith instance. A is a set of all attributes that has two parts, $A = C \cup D$ where C is a set of conditional attributes (features) and D is the decision attribute (classes). Mostly, datasets only have one decision attribute. V is a set of attributes domain, also, V_a is a subset of V such that consist of all the elements of attribute a. $f : U \times A \to V$ is an information function, which indicates the value of an instance in a specific feature. Some important concepts are defined as follows.

Definition 2.1. [15] *The first concept in rough set theory is indiscernibility. Indiscernibility relation is a function to find equivalence class of an instance with respect to a specific subset of attributes. Indiscernibility relation is defined as follows:*

$$IND(P) = \{(x, y) \in U^2 | \forall a \in P, f(x, a) = f(y, a)\}, \tag{1}$$

where $P \subseteq A$ is a subset of conditional attributes, x and y are instances of universe of discourse. A partition of U generated by $IND(P)$ is denoted by $U/IND(P)$ and is defined as:

$$U/(IND(P)) = \otimes\{a \in P : U/IND(\{a\})\}, \tag{2}$$

$$A \otimes B = \{X \cap Y : \forall X \in A, Y \in B, X \cap Y \neq \emptyset\}. \tag{3}$$

Definition 2.2. [15] *Let $IS = \langle U, A, V, f\rangle$ be an information system such that $X \subseteq U$ and $P \subseteq A$. The indiscernible instances of the universe of discourse which exactly belong to X with respect to $IND(P)$ are called the lower approximation of X with respect to $IND(P)$. Upper approximation of X with respect to $IND(P)$ is a set of indiscernible instances which probably belong to X with respect to $IND(P)$. Lower and upper approximations are denoted by $\underline{P}X$ and $\overline{P}X$ and are defined as:*

$$\underline{P}X = \{x \in U \mid [x]_P \subseteq X\}, \tag{4}$$

$$\overline{P}X = \{x \in U \mid [x]_P \cap X \neq \varnothing\}, \tag{5}$$

where $[x]_P$ is a set of instances which are indiscernible with x with respect to $IND(P)$. The ordered pair $\langle \underline{P}X, \overline{P}X\rangle$ is called rough set of X. Accuracy of approximation can be calculated by lower and upper approximations. Accuracy of approximation is denoted by $\alpha_P(X)$ and is defined as:

$$\alpha_P(X) = \frac{|\underline{P}X|}{|\overline{P}X|}. \tag{6}$$

Definition 2.3. [15] *Let $IS = \langle U, A, V, f\rangle$ be an information system and $A = C \cup D$ in which C is a set of conditional attributes (features) and D is the decision attribute (classes). Then, positive, negative and boundary regions are defined, respectively, as follows:*

$$POS_P(D) = \bigcup_{X \in U/D} \underline{P}X, \tag{7}$$

$$NEG_P(D) = U - \bigcup_{X \in U/D} \overline{P}X, \tag{8}$$

$$BND_P(D) = \bigcup_{X \in U/D} \overline{P}X - \bigcup_{X \in U/D} \underline{P}X, \tag{9}$$

where U/D is the partition of universe of discourse constructed by D. In fact, positive region is a partition of universe of discourse that is exactly classified in equivalence classes of U/D. Boundary region is a part of U that is possibility (not exactly) classified in classes of U/D. Negative region is a set of instances that is not classified correctly with respect to P. The most important concept on rough set theory is to find the dependency between two sets of attributes (usually between decision attributes and conditional attributes). This is the motivation for the next definition.

Definition 2.4. [15] *If P is a subset of conditional attributes and D is decision attribute, dependency between D and P is denoted by $\gamma_P(D)$ and is defined as follows:*

$$\gamma_P(D) = \frac{|POS_P(D)|}{|U|}. \tag{10}$$

2.2 Weighted rough set basic concepts

There are some methods to take action on class imbalanced data such as re-sampling, filtering inconsistent samples and weighted rough set [19, 20]. The weighted rough set is an extension of the rough set to deal with class imbalance data. It uses prior knowledge of instances which means that each sample has a weight. Let $WIS = \langle U, W, A, V, f\rangle$ be a weighted information system where U is a nonempty set of samples, A is the feature set and W is the weight of samples. In the weighted rough set, the most concepts and definitions, such as the definitions of lower and upper approximations, positive, negative and boundary regions, are the same as the classical rough set. However, the accuracy of approximation and dependency degree are different. The weighted accuracy of approximation is formalized as follows:

$$\alpha_P^W(X) = \frac{|\underline{P}X|_W}{|\overline{P}X|_W}, \tag{11}$$

where $|\underline{P}X|_W = \sum_{x_i \in \underline{P}X}(w_i)$ and $|\overline{P}X|_W = \sum_{x_i \in \overline{P}X}(w_i)$ are the weighted cardinality of lower and upper approximations, respectively. The weighted dependency degree is represented as:

$$\gamma_P^W(D) = \frac{|POS_P(D)|_W}{|U|_W}, \tag{12}$$

where P is a subset of conditional attributes, D is the decision attribute, $|POS_P(D)|_W = \sum_{x_i \in POS_P(D)}(w_i)$ and $|U|_W = \sum_{x_i \in U}(w_i)$.

2.3 Fuzzy rough set basic concepts

Fuzzy rough set [15] is a powerful extension of the rough set to deal with vagueness and uncertainty in continuous datasets. This extension makes use of fuzzy relations to explain indiscernibility and discernibility concepts. Assume that U is a non-empty universe of discourse and R is a fuzzy relation on U such that R satisfies the following properties [15, 9]:

1. Reflexivity: $R(x, x) = 1, \forall x \in U$.
2. Symmetry: $R(x, y) = R(y, x), \forall x, y \in U$.
3. Transitivity: $R(x, z) \geq min_y(R(x, y), R(y, z))$.

The fuzzy similarity function used to calculate the equivalence relation is defined as follows [15]:

$$R = 1 - \frac{|a(x) - a(y)|}{|a_{\max} - a_{\min}|}, \tag{13}$$

where a is an attribute, x and y are two instances and a_{max} and a_{min} are the maximum and minimum values in the attribute a, respectively.

Definition 2.5. [15] *Let U be universe of discourse, R be a fuzzy relation on U and $F(U)$ be a fuzzy power set of U where F is a fuzzy concept to be approximated. Then the fuzzy lower and upper approximations are defined as follows:*

$$\underline{R}F(x) = \inf_{y \in U} \max\{1 - R(x, y), F(y)\}, \tag{14}$$

$$\overline{R}F(x) = \sup_{y \in U} \min\{R(x, y), F(y)\}. \tag{15}$$

where $F(y)$ is the membership of y belonging to fuzzy set F.

The pair of $\langle \underline{R}F(x), \overline{R}F(x)\rangle$ is called a fuzzy rough set. Fuzzy lower approximation ensures that a sample belongs to a class and the fuzzy upper approximation shows its possibility. In order to detect relevant features by means of the fuzzy rough set, a function should be defined in such a way as to show a sample exactly belongs to a class. This leads to the following concepts.

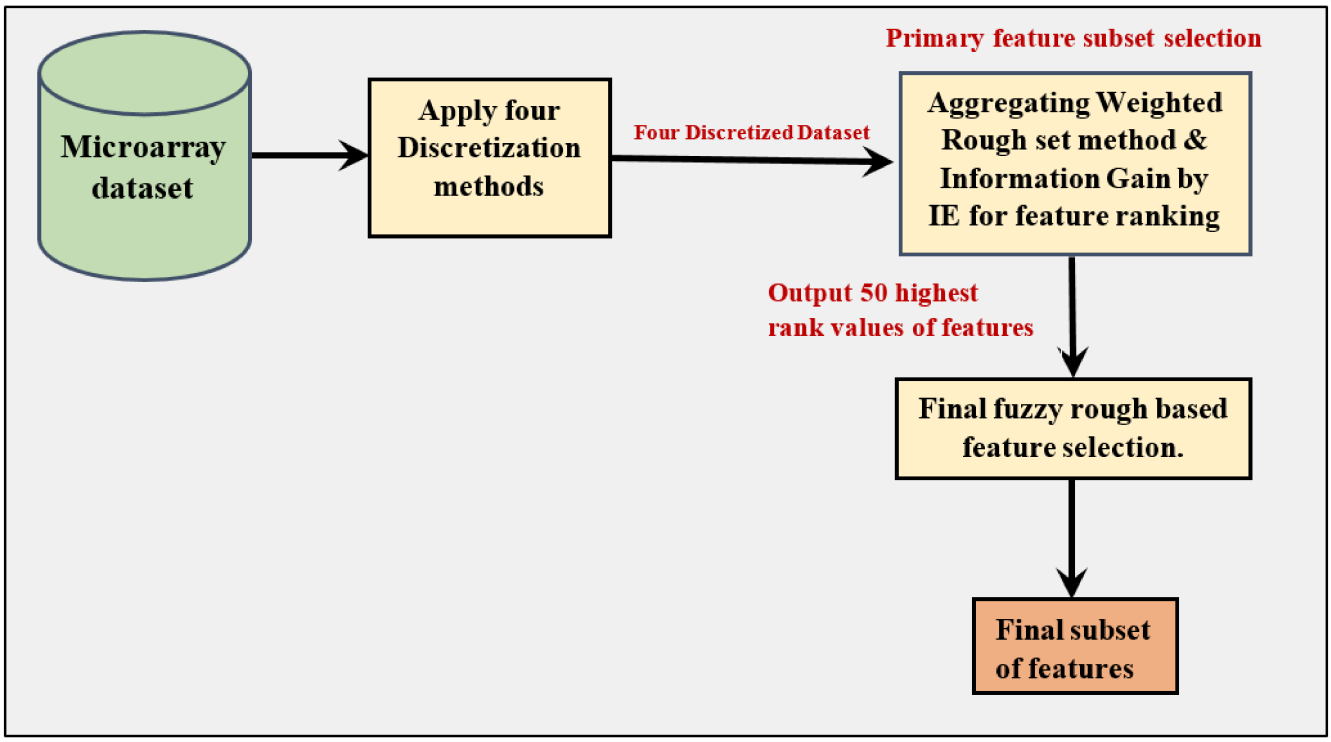

Figure 1: The framework of proposed method

Definition 2.6. [15] *Let U be the universe and $F(U)$ be a fuzzy power set of U. Fuzzy positive region and fuzzy dependency degree with respect to the features set P are defined, respectively, as follows:*

$$\mu_{POS_P(D)}(x) = \sup_{F \in U/D} (\underline{P}F(x)), \tag{16}$$

$$\widetilde{\gamma}_P(D) = \frac{\sum_{x \in U} \mu_{POS_P(D)}(x)}{|U|}. \tag{17}$$

The fuzzy dependency denoted by $\widetilde{\gamma}_P(D)$ indicates the relevancy of a subset of features. Considering that B is a subset of P, a measure of the significance of B is defined as follows:

$$\widetilde{\sigma}_{P,D}(B) = \frac{\widetilde{\gamma}_P(D) - \widetilde{\gamma}_{P-B}(D)}{\widetilde{\gamma}_P(D)} = 1 - \frac{\widetilde{\gamma}_{P-B}(D)}{\widetilde{\gamma}_P(D)} \tag{18}$$

Note that the case $\widetilde{\sigma}_{P,D}(B) = 0$ reveals that the subset B is not important and it can be deleted from set P.

2.4 Information gain

Information gain [13, 34] assigns a rank to each feature to evaluate discrimination capability of each feature. At first, entropy should be calculated as follows:

$$Entropy(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-, \tag{19}$$

where p_+ and p_- are probability of positive class and probability of negative class, respectively. Then, information gain is calculated as follows:

$$Gain(S, F_i) = Entropy(S) - \sum_{v \subset F_i} \frac{|S_v|}{|S|} Entropy(S_v), \tag{20}$$

where F_i is the ith feature and S_v is a subset of instances in F_i that have value v and S is the set of all instances.

3 Proposed method

In this section, a feature selection method for binary and class imbalance microarray datasets is presented. The proposed method applies four discretization methods [7, 25] to discretize continuous dataset and four discrete datasets are generated. Then, it uses a rough set method and mutual information [9] to compute four dependency degrees and information gains by discrete datasets for each feature. Finally, the final rank-value is computed by an aggregation method [32]. The proposed method consists of two steps. In the first step, weighted rough dependency degree and information gain are utilized to select a primary set of all conditional features. Then, in the second step, the fuzzy rough set method is used to find the most important feature in the primary subset by a measure of the significance of features. Finally, the second step removes all insignificant and redundant features from the primary subset and the final subset of relevant features is made by this step. The framework of the proposed method is shown in Figure 1. The discretization approaches being applied in the suggested method are called equal width, equal frequency, global equal width [25] and MDLP (Minimum Description Length Principle) [12]. The equal width and the equal frequency employ the equal frequencies or equal width binning algorithm. These methods discretize each random variable (each feature) of the data into n bins. The global equal width discretizes the range of the random vector data into n bins. The MDLP discretizes the continuous attributes of data matrix using entropy criterion with the Minimum Description Length as stopping rule.

3.1 Sample weighting approach

In the rough set, the samples in the positive region are correctly classified and sample weighting approach is used to make a pressure for sensitive samples in order to put them in the positive region. Sensitive samples for classification are the minor class samples in the class boarder in class imbalance data. Each sample weight is computed as follows:

$$w_i = \frac{C_j}{d(x_i, Avg_j)}, \quad i = 1, 2, \ldots, n \quad j = 1, \ldots, c \tag{21}$$

$$Avg_j = \frac{1}{n_j} \sum_{k=1}^{n_j} x_k, \tag{22}$$

where w_i is weight for ith sample, n and c are number of samples and classes. $d(x, y)$ is a distance function between x and y, C_j is a constant weight of jth class, Avg_j is the central sample in jth class and n_j is the number of samples in class j.

3.2 Aggregation measure

Information Energy (IE) is an aggregation measure based on hesitant fuzzy set concepts [32, 4]. Let U be a universe of discourse. Hesitant fuzzy set A is defined on U by, $A = \{\langle x, h_A(x)\rangle | x \in U\}$, (23)

where $h_A(x)$ that called hesitant fuzzy element of x , is the set of all possible values in the interval $[0, 1]$. For a hesitant fuzzy set, IE is defined as follows. $E_{HFS}(A) = \sum_{i=1}^{n} \left(\frac{1}{l_i} \sum_{j=1}^{l_i} h^2_{A\sigma(j)}(x_i) \right)$, (24)

where n is the cardinality of the universe, l_i is the number of membership values and $h_{A\sigma(j)}$ is the element of ith hesitant fuzzy set.

3.3 Primary feature subset selection

In this step, a rank-value is computed for each feature by using discretization methods, weighted rough set and mutual information theory. At first, continuous datasets are discretized by employing four discretization methods. Then, for each feature four weighted dependency degrees and four information gains are computed according to corresponding discretization methods. Then, the dependency degrees and information gains are aggregated by IE. The IE values are considered as the rank-values. Finally, a predefined number of features with higher rank-value are selected as the primary subset, in this paper the predefined number of features is considered to be 50. Pseudo-code of primary feature subset selection is shown in Algorithm 3.1.

The aggregation of weighted rough set and information gain is performed by Algorithm 3.1. In Algorithm 3.1 m discretized datasets and weights of samples being computed by Eq 3.1 are used as the inputs of algorithm. Each

Algorithm 3.1 Primary feature subset selection

Input : discretized datasets and weight of samples.
Output : a subset of features (PS).

for $i = 1$ to d(d is the number of features) **do**
  $Q_i \leftarrow \{\}$
  for $j = 1$ to m(m is the number of discretized datasets) **do**
    $WDD(i,j) \leftarrow \gamma^W_{i,j}(D)$, Eq 12.
    $IG(i,j) \leftarrow Gain(S, F_{i,j})$, Eq 20.
  end for
  $Q_i \leftarrow Q_i \cup \{WDD_i, IG_i\}$
  $IE(i) \leftarrow E_{HFS}(Q_i)$, Eq 24
end for
Finally, features with highest IE values are selected as primary subset(PS).

discretized dataset represents a different source of information and knowledge. Finally, the primary subset selection algorithm is depicted as a flowchart in Figure 2. In the flowchart as the Algorithm 3.1 , i is the index of features and j is the jth discretized form of feature i with respect to jth discretization method. At the first time, $i = 1$ and $j = 1$ and vector $Q_i = \{\}$. Q_i stands for the hesitant fuzzy set that maintains the hesitant fuzzy elements for ith feature. These hesitant elements are the values of $WDD_{i,j}$ and $IG_{i,j}$ being computed by Eq 11 and Eq 20.

3.4 Fuzzy rough-based elimination

In this step, a significant measure is utilized to eliminate the irrelevant and redundant features from the primary subset of features. In algorithm 1, a primary subset of features is selected and then these features are fed into the second step called Algorithm 2. As it can be seen in Algorithm 3.2, fuzzy dependency degrees of primary features are computed. Then a feature is selected as the candidate of elimination, it is the redundant feature that should be deleted from the primary subset. New dependency degree without this feature is computed. Then, the fuzzy measure of the significance of the candidate feature is computed. If its value is zero then, candidate feature is a redundant feature and is deleted from the primary set. The significant measure is computed for all features in the primary subset. Finally, the final subset of features is selected. Final fuzzy rough-based feature selection step is shown in Algorithm 3.2.

Algorithm 3.2 Fuzzy rough-based elimination algorithm

Input : primary subset of features (PS) and original dataset.
Output : final selected subset of features (FS).

$\widetilde{\gamma}_{PS}(D) = \frac{\sum_{x\in U} \mu_{POS_{PS^{(D)}}}(x)}{|U|}$, Eq 17.
$FS \leftarrow PS$
for $i = 1$ to z(z is cardinality of (PS)) **do**
  $\widetilde{\gamma}_{FS-\{a_i\}}(D) = \frac{\sum_{x\in U} \mu_{POS_{FS-\{a_i\}^{(D)}}}(x)}{|U|}$, Eq 17.
  $\widetilde{\sigma}_{FS,D}(a_i) = 1 - \frac{\widetilde{\gamma}_{FS-\{a_i\}}(D)}{\widetilde{\gamma}_{PS}(D)}$, Eq 18.
  if $\widetilde{\sigma}_{FS,D}(a_i) = 0$ **then**
    $FS \leftarrow FS - \{a_i\}$.
  end if
end for

The fuzzy rough-based elimination algorithm starts from the primary subset. In the flow of algorithm 3.2 the redundant features in the primary subset are detected and eliminated. According to algorithm 3.2, fuzzy dependency degree of the primary subset is computed by Eq 2.6. Then feature a is selected as candidate feature and fuzzy dependency degree of the primary subset without candidate feature is computed and is compared with fuzzy dependency degree of the primary subset with a. If these two degrees are equal, then candidate feature a is redundant feature and should be deleted. This process repeats for all features in the primary subset and all redundant features are deleted. The Figure 3 explains the step of algorithm 3.2.

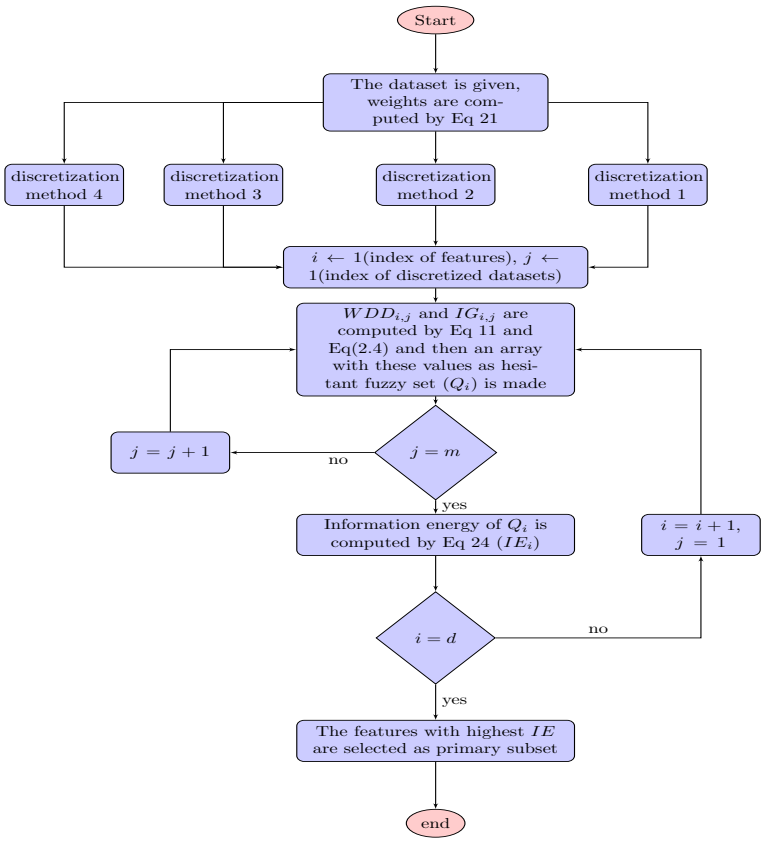

Figure 2: The flowchart of primary subset selection stage

4 Experimental study

In this section, the experimental results are presented and analyzed to show the accuracy and performance of the proposed method. In the following, microarray datasets utilized in this experiment, are presented and then the cross validation and evaluation measures are introduced and finally, the results of the proposed method are shown.

4.1 Microarray datasets

As we said, the microarray datasets that utilized in this paper are high dimensional data with class imbalance instances. The proposed method is experimented across 7 microarray datasets that described in Table 1 [2, 31]. The columns of the table are the number of features (#Feats), the number of samples (#Samp), the distribution of different classes (minor and major class) and imbalance ratio of the dataset (the division of major samples on minor samples).

Dataset	#Feats	#Samp	(%min,%maj)	IR
Brain	12625	21	(33.33,66.67)	1.71
CNS	7129	60	(35.00,65.00)	1.86
Colon	2000	62	(35.48,64.52)	1.82
DLBCL	4026	47	(48.94,51.06)	1.04
GLI	22283	85	(30.59,69.41)	2.27
Ovarian	15154	253	(35.97,64.03)	1.78
SMK	19993	187	(48.13,51.87)	1.08

Table 1: The description of microarray datasets which are utilized in this paper

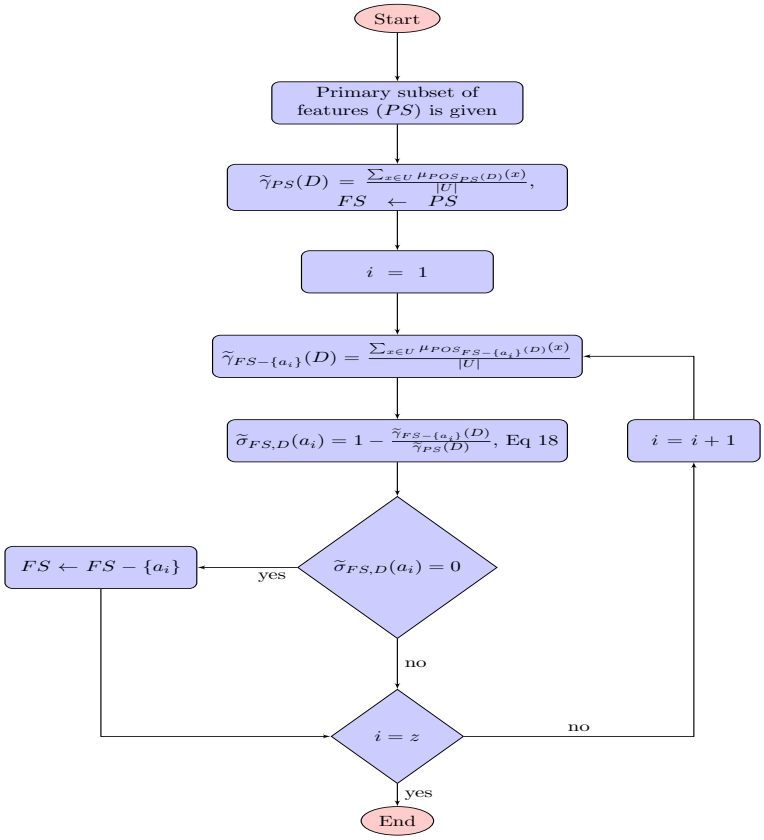

Figure 3: The flowchart of Fuzzy rough-based elimination algorithm

4.2 Validation approach

The proposed method and the selected set of features should be evaluated. There are many techniques for evaluating a method, this paper utilizes Distribution Optimally Balanced Stratified Cross-Validation (DOB-SCV) [27]. In this technique, a sample is selected randomly and $k-1$ (k is the number of folds) nearest samples of the selected sample with same class are found and are arranged in different sets. This process is repeated until all unassigned samples to be assigned to a set and finally, k sets are built. Eventually, k folds (k subsets) are achieved and one of them is used as the test set. The process continues until each fold to be selected as the test set.

4.3 Evaluation measures

This paper utilizes three well-known classifiers from different types so as to evaluate the selected set of features and proposed feature selection method. Selected classifiers in this paper are C4.5 which is a rule based classifier and the C4.5 is an extension of ID3 algorithm that generate decision trees also it is often referred to as a statistical classifier, Naive Bayes which is a probabilistic classifier based on applying Bayes theorem with strong independence assumptions between the features and Support Vector Machine (SVM) which is a supervised learning model with associated learning algorithms and The SVM is optimization based classifier. Whereas, three mentioned classifiers widely are used in practice.

Furthermore, this paper utilizes evaluation metrics that designed for binary classification (positive and negative class) problems such as Sensitivity, Specificity, G-mean and Accuracy values [2]. Sensitivity measure shows how well the positive class samples are predicted in test step and specificity shows how well the negative class examples are predicted and finally, accuracy measure shows how well all the test samples are predicted in a correct category. G-mean is the geometric mean of sensitivity and specificity. The high value of G-mean means the high values of sensitivity and specificity simultaneously. These measures are computed by true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

$$Sensitivity = \frac{TP}{TP + FN} \tag{25}$$

$$Specificity = \frac{TN}{TN + FP} \tag{26}$$

$$G - mean = \sqrt{Sensitivity \times Specificity} \tag{27}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{28}$$

4.4 Results and discussion

In this section, the numeric results of the proposed method and other methods are illustrated in different kind of tables and charts. The number of selected features by the proposed method and the other methods in literature are shown in Table 2. The number of selected features is the average of selected features in DOB-SCV cross validation procedure with 5 folds. The proposed method removes more than 90% of features in datasets. In the following, the results of different classifiers in different evaluation measures are illustrated by a variety of tables and figures.

According to Tables 3, 4 and 5, the results of different classifiers on microarray datasets in aforementioned evaluation measures are illustrated. Obviously, the presented method has a significant difference with other methods in term of classification performance.

Methods	Brain	CNS	Colon	DLBCL	GLI	Ovarian	SMK
ReliefF-10	10	10	10	10	10	10	10
ReliefF-50	50	50	50	50	50	50	50
mRMR-10	10	10	10	10	10	10	10
mRMR-50	50	50	50	50	50	50	50
MRMR-HFS	27.2	5.2	5.6	2.2	8.4	4.4	6.2
RREFS	20.2	48	49.6	37.6	68	202.4	149.6
RSMRMSI(crisp discretization)	1	1	1	1	1	1	1
RSMRMSII(fuzzy discretization)	1	1	1	1	1	1	1
FEPM	1	1	1	1	1	1	1
RHEPM-MRMDMS	1	10.6	9.2	3.2	8.8	13.6	40.4
Proposed method	16.2	41	39.6	35.2	41.2	41.6	46.8

Table 2: The number of selected feature in microarray datasets

The average performance of three classifiers in terms of Sensitivity, Specificity, G-mean and Accuracy are shown in Figure 4, it shows the proposed method has an appropriate performance and it has better results than the other methods. Figure 5 has the same meaning as Figure 4, whereas, the vertical axis of Figure 4 consist of feature selection methods and the vertical axis of Figure 5 is based on four evaluation measures.

Figure 6 shows that the proposed method has the better result in evaluation measures based on C4.5 classifier. The results of each classifier are shown in different radar charts. The proposed method has significant different and it has appropriate results across all evaluation measures, the results of Naive Bayes classifier are presented in Figure 7.

This chart shows proposed method achieves the best results by Naive Bayes classifier and it has significant difference with other two methods.

SVM classifier results are same as Naive Bayes classifier, the proposed method has appropriate results in all evaluation measures for this classifier. The results of SVM are shown in Figure 8.

Finally, the superiority of the proposed method to other methods are evaluated by the statistical non-parametric Wilcoxon Signed Ranked test and the results of this test are shown in Table 6 [30]. All methods in this Table are

Methods		Measures	Brain	CNS	Colon	DLBCL	GLI-85	Ovarian	SMK	Avg
ReliefF	#10	Ac	0.48	0.55	0.77	0.84	0.81	0.96	0.66	0.72
		Se	0.1	0.76	0.69	0.82	0.91	0.93	0.67	0.7
		Sp	0.63	0.15	0.82	0.87	0.57	0.98	0.64	0.67
		G-mn	0.25	0.34	0.75	0.84	0.72	0.95	0.65	0.65
	#50	Ac	0.53	0.56	0.76	0.8	0.85	0.98	0.68	0.74
		Se	0.6	0.69	0.63	0.82	0.86	0.98	0.75	0.76
		Sp	0.5	0.34	0.82	0.79	0.81	0.98	0.6	0.69
		G-mn	0.55	0.48	0.72	0.8	0.83	0.98	0.67	0.72
mRMR	#10	Ac	0.8	0.6	0.79	0.78	0.8	0.98	0.68	0.78
		Se	0.8	0.71	0.65	0.78	0.79	0.98	0.75	0.78
		Sp	0.83	0.38	0.87	0.8	0.81	0.98	0.61	0.75
		G-mn	0.81	0.52	0.75	0.8	0.8	0.98	0.68	0.76
	#50	Ac	0.84	0.6	0.82	0.76	0.78	0.98	0.71	0.78
		Se	0.9	0.66	0.69	0.78	0.76	0.98	0.71	0.78
		Sp	0.83	0.47	0.9	0.74	0.82	0.98	0.7	0.78
		G-mn	0.86	0.56	0.79	0.76	0.79	0.98	0.7	0.78
MRMR-HFS		Ac	0.93	0.62	0.83	0.81	0.85	0.96	0.66	0.81
		Se	1	0.67	0.9	0.83	0.65	0.96	0.7	0.82
		Sp	0.9	0.53	0.73	0.78	0.93	0.96	0.63	0.78
		G-mn	0.95	0.6	0.81	0.8	0.78	0.96	0.66	0.79
RREFS		Ac	1	0.91	0.91	0.86	0.89	0.96	0.73	0.86
		Se	1	0.69	0.78	0.81	0.93	0.98	0.84	0.89
		Sp	1	0.79	0.83	0.87	0.98	0.97	0.76	0.89
		G-mn	1	0.88	0.9	0.84	0.95	0.98	0.8	0.89
RSMRMSI		Ac	0.89	0.58	0.65	0.73	0.68	0.9	0.52	0.71
		Se	0.93	0.71	0.75	0.72	0.8	0.89	0	0.69
		Sp	0.8	0.34	0.43	0.74	0.63	0.9	1	0.69
		G-mn	0.76	0.31	0.36	0.7	0.69	0.89	0	0.53
RSMRMSII		Ac	0.72	0.62	0.71	0.75	0.62	0.83	0.52	0.68
		Se	0.87	0.88	0.93	0.8	0.69	0.85	0	0.72
		Sp	0.5	0.15	0.31	0.7	0.59	0.83	1	0.58
		G-mn	0.5	0.11	0.38	0.74	0.62	0.83	0	0.45
FEPM		Ac	0.67	0.57	0.82	0.89	0.74	0.96	0.59	0.75
		Se	1	0.64	0.88	0.91	0.67	0.95	0.67	0.81
		Sp	0	0.45	0.73	0.87	0.78	0.98	0.52	0.62
		G-mn	0	0.42	0.8	0.88	0.7	0.96	0.57	0.62
RHEPM-MRMDMS		Ac	1	0.56	0.87	0.83	0.86	0.98	0.65	0.82
		Se	1	0.64	0.93	1	0.69	0.96	0.62	0.83
		Sp	1	0.42	0.79	0.66	0.93	0.99	0.67	0.78
		G-mn	1	0.48	0.84	0.79	0.79	0.97	0.64	0.79
Proposed Method		Ac	1	0.89	0.95	0.98	0.94	0.99	0.92	0.95
		Se	1	0.95	0.95	0.96	0.83	0.98	0.92	0.94
		Sp	1	0.79	0.96	1	1	0.99	0.91	0.95
		G-mn	1	0.85	0.95	0.98	0.88	0.99	0.91	0.94

Table 3: Experimental results for C4.5 classifier on microarray datasets after performing DOB-SCV validation with 5 folds

compared with the proposed method in terms of G-mean. Each method that have a $p-value \leq 0.05$ is rejected by the proposed method. Table 6 shows that the null hypothesis is rejected for all comparisons and subsequently the proposed method is significantly better than the other methods.

Methods		Measures	Brain	CNS	Colon	DLBCL	GLI-85	Ovarian	SMK	Avg
ReliefF	#10	Ac	0.26	0.65	0.84	0.93	0.84	0.96	0.67	0.74
		Se	0.3	0.69	0.72	0.96	0.88	0.95	0.71	0.74
		Sp	0.2	0.57	0.9	0.91	0.74	0.96	0.62	0.7
		G-mn	0.24	0.61	0.8	0.93	0.81	0.95	0.66	0.72
	#50	Ac	0.21	0.67	0.84	0.96	0.86	0.98	0.68	0.74
		Se	0.3	0.71	0.72	0.96	0.86	0.95	0.77	0.75
		Sp	0.13	0.58	0.9	0.96	0.85	0.99	0.59	0.71
		G-mn	0.2	0.64	0.8	0.96	0.85	0.97	0.67	0.73
mRMR	#10	Ac	0.75	0.68	0.82	0.98	0.87	0.99	0.71	0.82
		Se	0.8	0.74	0.77	1	0.93	0.98	0.74	0.85
		Sp	0.73	0.58	0.85	0.96	0.73	0.99	0.66	0.79
		G-mn	0.76	0.66	0.81	0.98	0.82	0.98	0.7	0.82
	#50	Ac	0.77	0.67	0.79	0.98	0.85	0.98	0.67	0.82
		Se	0.4	0.71	0.82	0.96	0.87	0.96	0.7	0.77
		Sp	1	0.59	0.77	1	0.77	0.99	0.64	0.82
		G-mn	0.63	0.65	0.79	0.98	0.82	0.97	0.67	0.79
MRMR-HFS		Ac	0.96	0.67	0.84	0.95	0.90	0.98	0.70	0.86
		Se	0.90	0.80	0.88	0.91	0.89	0.96	0.65	0.86
		Sp	1.00	0.43	0.77	1.00	0.90	0.99	0.72	0.83
		G-mn	0.95	0.58	0.82	0.95	0.90	0.97	0.68	0.84
RREFS		Ac	1.00	0.72	0.79	0.84	0.97	0.97	0.77	0.88
		Se	1.00	0.69	0.78	0.81	0.93	0.98	0.83	0.89
		Sp	1.00	0.79	0.83	0.87	0.98	0.97	0.76	0.89
		G-mn	1.00	0.68	0.80	0.84	0.95	0.98	0.79	0.89
RSMRMSI		Ac	0.80	0.60	0.60	0.75	0.71	0.91	0.52	0.70
		Se	1.00	0.76	0.45	0.76	0.84	0.91	0.00	0.67
		Sp	0.40	0.29	0.88	0.74	0.65	0.91	1.00	0.69
		G-mn	0.40	0.31	0.59	0.72	0.73	0.91	0.00	0.52
RSMRMSII		Ac	0.63	0.62	0.63	0.75	0.62	0.84	0.52	0.66
		Se	0.87	0.88	0.48	0.80	0.76	0.86	0.00	0.66
		Sp	0.20	0.15	0.92	0.70	0.56	0.83	1.00	0.62
		G-mn	0.16	0.11	0.62	0.74	0.63	0.84	0.00	0.44
FEPM		Ac	0.67	0.55	0.82	0.89	0.73	0.96	0.59	0.75
		Se	1.00	0.73	0.88	0.91	0.63	0.95	0.67	0.82
		Sp	0.00	0.25	0.73	0.87	0.78	0.98	0.52	0.59
		G-mn	0.00	0.27	0.80	0.88	0.68	0.96	0.57	0.60
RHEPM-MRMDMS		Ac	0.96	0.62	0.86	0.88	0.87	0.98	0.67	0.83
		Se	0.93	0.71	0.88	0.92	0.77	0.96	0.59	0.82
		Sp	1.00	0.44	0.83	0.84	0.92	1.00	0.74	0.82
		G-mn	0.96	0.54	0.84	0.87	0.83	0.98	0.66	0.81
Proposed Method		Ac	1.00	0.91	0.94	1.00	0.94	0.98	0.91	0.95
		Se	1.00	0.95	0.93	1.00	0.83	0.95	0.89	0.93
		Sp	1.00	0.83	0.96	1.00	1.00	1.00	0.92	0.96
		G-mn	1.00	0.88	0.94	1.00	0.88	0.97	0.90	0.94

Table 4: Experimental results for Naive Bayes classifier on microarray datasets after performing DOB-SCV validation with 5 folds

Methods	R^+	R^-	Exact P-value	Asymptotic P-value
ReliefF#10	28.0	0.0	0.015626	0.014248
ReliefF#50	28.0	0.0	0.015626	0.014248
mRMR#10	21.0	0.0	0.03126	0.021098
mRMR#50	21.0	0.0	0.03126	0.021098
MRMR-HFS	28.0	0.0	0.015626	0.014248
RREFS	24.0	4.0	0.10938	0.069811
RSMRMSI	28.0	0.0	0.015626	0.014248
RSMRMSII	28.0	0.0	0.015626	0.014248
FEPM	28.0	0.0	0.015626	0.014248
RHEPM-MRMDMS	21.0	0.0	0.03126	0.021098

Table 6: Results obtained by the Wilcoxon test for comparing the proposed method with the other algorithms in terms of G-mean.

Methods		Measures	Brain	CNS	Colon	DLBCL	GLI-85	Ovarian	SMK	Avg
ReliefF	#10	Ac	0.45	0.63	0.82	0.95	0.84	0.98	0.65	0.76
		Se	0.1	0.84	0.58	1	0.93	0.93	0.75	0.73
		Sp	0.6	0.25	0.95	0.91	0.62	1	0.54	0.7
		G-mn	0.24	0.46	0.74	0.95	0.76	0.96	0.64	0.68
	#50	Ac	0.64	0.63	0.84	0.94	0.88	0.99	0.72	0.81
		Se	0.3	0.74	0.81	1	0.95	0.98	0.83	0.8
		Sp	0.8	0.43	0.85	0.88	0.73	1	0.58	0.75
		G-mn	0.49	0.56	0.83	0.94	0.83	0.99	0.69	0.76
mRMR	#10	Ac	0.45	0.65	0.82	0.98	0.87	1	0.71	0.78
		Se	0.2	0.87	0.72	1	0.95	1	0.74	0.78
		Sp	0.6	0.24	0.87	0.96	0.69	1	0.66	0.72
		G-mn	0.35	0.46	0.79	0.98	0.81	1	0.7	0.73
	#50	Ac	0.58	0.7	0.84	0.93	0.87	1	0.66	0.8
		Se	0.1	0.74	0.78	0.95	0.95	1	0.73	0.75
		Sp	0.8	0.63	0.87	0.92	0.7	1	0.57	0.78
		G-mn	0.28	0.68	0.82	0.93	0.82	1	0.65	0.74
MRMR-HFS		Ac	0.58	0.65	0.87	0.94	0.92	0.99	0.75	0.83
		Se	1.00	0.71	0.93	0.87	0.81	0.99	0.72	0.86
		Sp	0.40	0.53	0.78	1.00	0.97	1.00	0.76	0.79
		G-mn	0.63	0.62	0.84	0.93	0.88	0.99	0.75	0.81
RREFS		Ac	0.96	0.92	0.92	0.94	0.94	0.98	0.77	0.88
		Se	0.93	0.95	0.95	0.87	0.88	0.97	0.66	0.87
		Sp	1.00	0.88	0.92	1.00	0.97	0.97	0.88	0.84
		G-mn	0.96	0.91	0.80	0.93	0.92	0.97	0.76	0.84
RSMRMSI		Ac	0.72	0.65	0.64	0.75	0.74	0.90	0.52	0.70
		Se	0.87	1.00	0.85	0.74	0.55	0.80	0.00	0.69
		Sp	0.40	0.00	0.27	0.75	0.83	0.96	1.00	0.60
		G-mn	0.32	0.00	0.35	0.72	0.65	0.87	0.00	0.41
RSMRMSII		Ac	0.63	0.65	0.69	0.70	0.76	0.83	0.52	0.68
		Se	0.93	1.00	0.83	0.68	0.53	0.73	0.00	0.67
		Sp	0.00	0.00	0.43	0.73	0.86	0.89	1.00	0.56
		G-mn	0.00	0.00	0.51	0.69	0.59	0.80	0.00	0.37
FEPM		Ac	0.63	0.62	0.82	0.87	0.74	0.96	0.61	0.75
		Se	0.80	0.74	0.90	0.87	0.55	0.92	0.61	0.77
		Sp	0.20	0.39	0.68	0.86	0.83	0.99	0.61	0.65
		G-mn	0.14	0.52	0.78	0.86	0.65	0.96	0.60	0.64
RHEPM-MRMDMS		Ac	0.87	0.57	0.77	0.89	0.87	1.00	0.66	0.80
		Se	0.80	0.61	0.83	1.00	0.73	0.99	0.60	0.79
		Sp	1.00	0.48	0.70	0.78	0.93	1.00	0.72	0.80
		G-mn	0.89	0.52	0.73	0.88	0.81	0.99	0.66	0.78
Proposed Method		Ac	0.88	0.92	0.97	1.00	0.86	1.00	0.80	0.92
		Se	0.87	0.98	0.98	1.00	0.82	0.99	0.78	0.91
		Sp	0.90	0.84	0.96	1.00	0.88	1.00	0.82	0.91
		G-mn	0.88	0.88	0.97	1.00	0.84	0.99	0.80	0.91

Table 5: Experimental results for SVM classifier on microarray datasets after performing DOB-SCV validation with 5 folds

5 Conclusion

In this paper a hybrid soft computing approach was proposed in which the hesitant fuzzy set, the rough set concepts and discretization methods were used to select the best subset of features in microarray class imbalanced datasets. Experimental results were conducted as well as statistical comparisons by Wilcoxon Signed Ranked test. The results confirmed the superiority of our proposed approach in comparison with some recently developed feature selection methods. The proposed method has the following advantages:

- Some soft computing approaches have been efficiently combined via hesitant fuzzy sets for feature selection.
- The time complexity of the proposed method is linear with respect to the number of features and therefore it can be easily extended to high dimensional datasets. The order of presented method is $O(d)$ in which d is the total number of features.

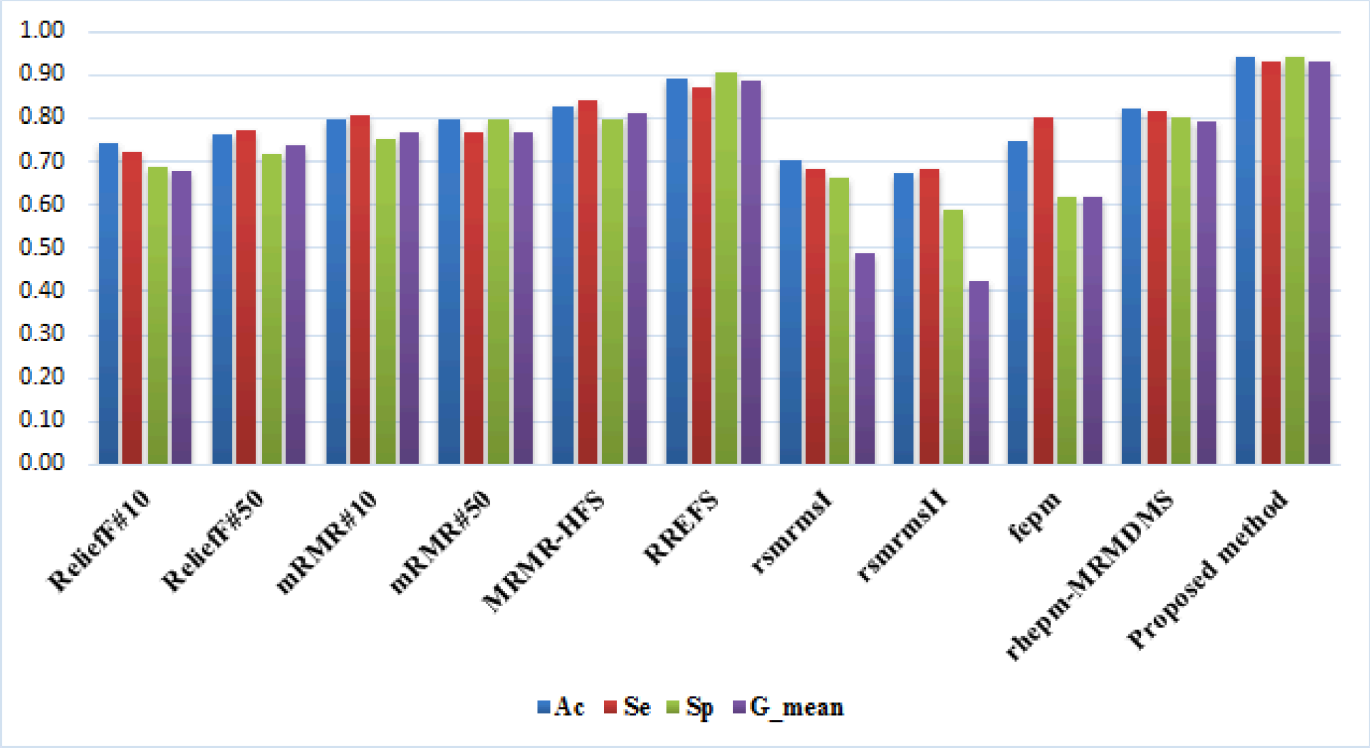

Figure 4: Average performance of three classifiers for different feature selection methods.

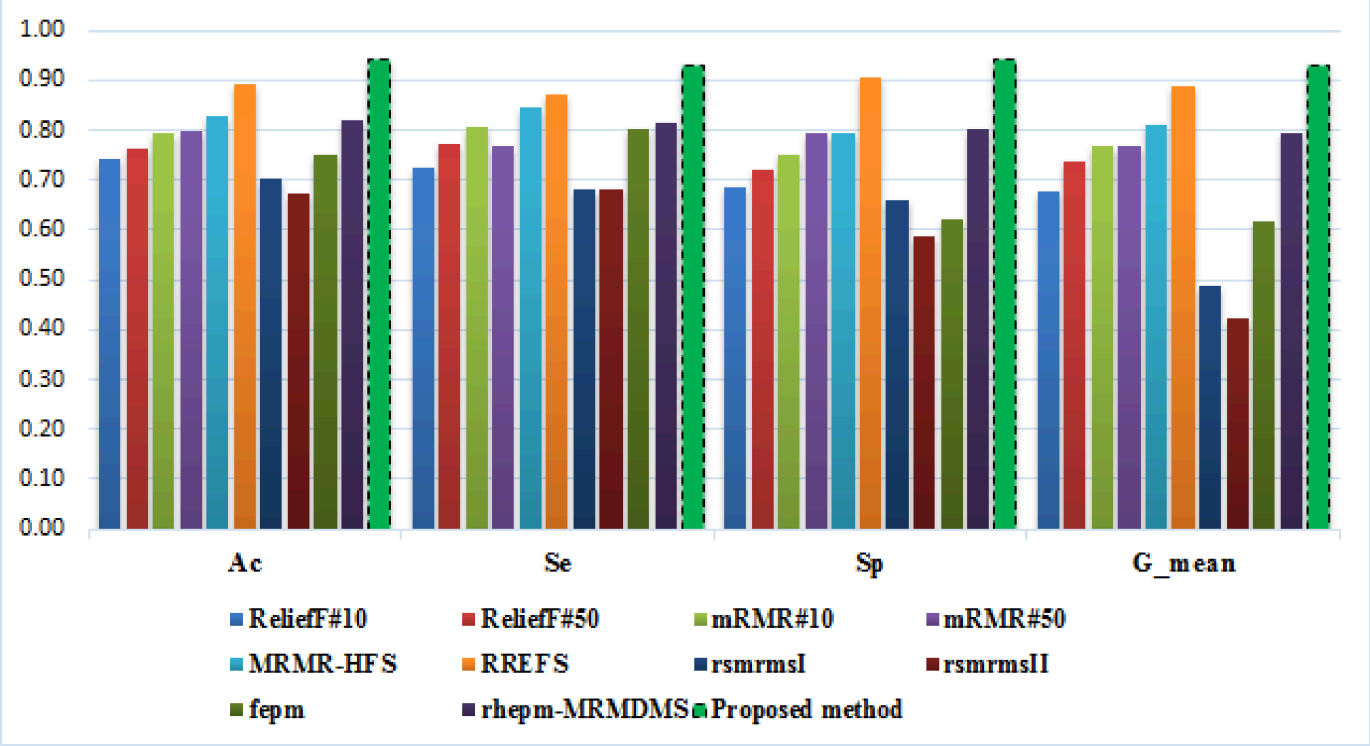

Figure 5: Average performance of three classifiers based on different evaluation measures.

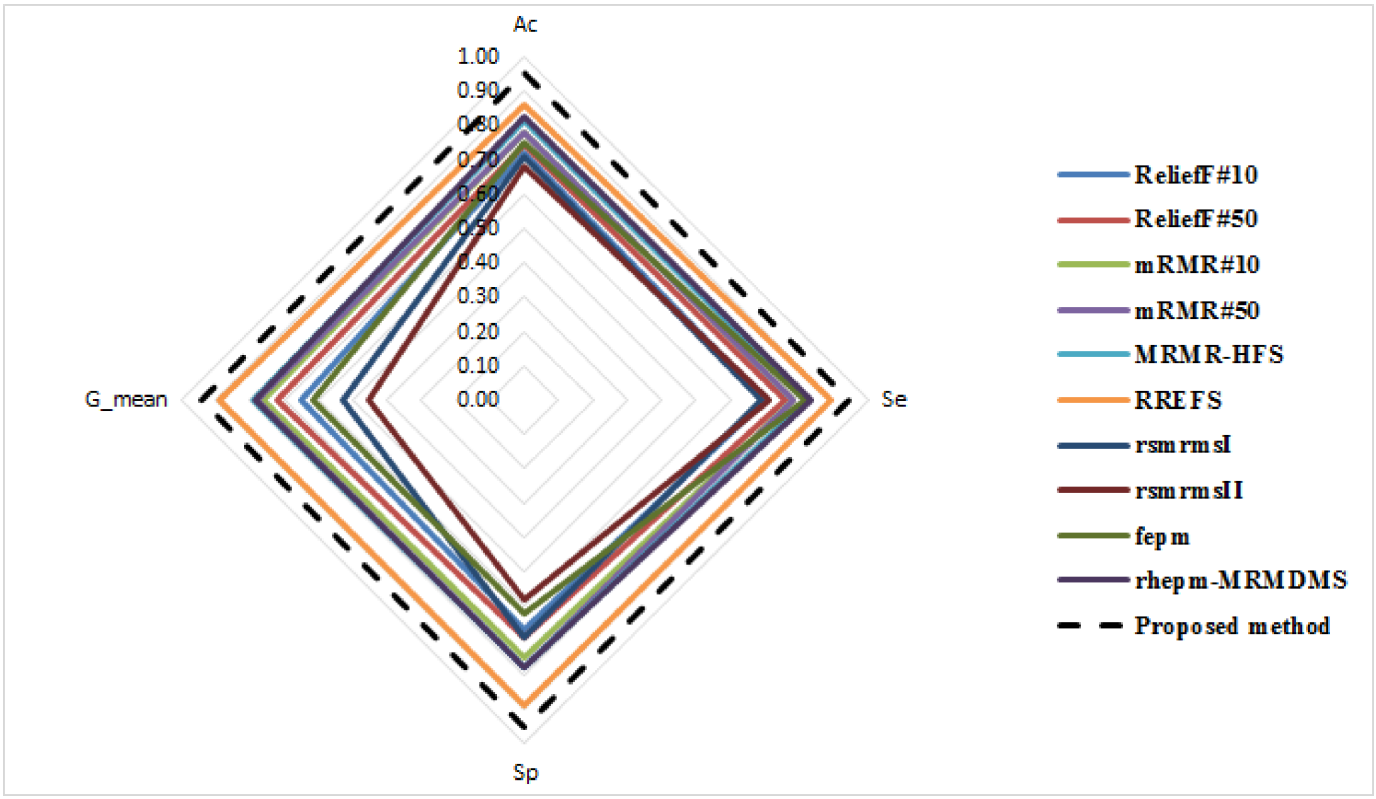

Figure 6: Radar chart of C4.5 classifier results based on evaluation measures.

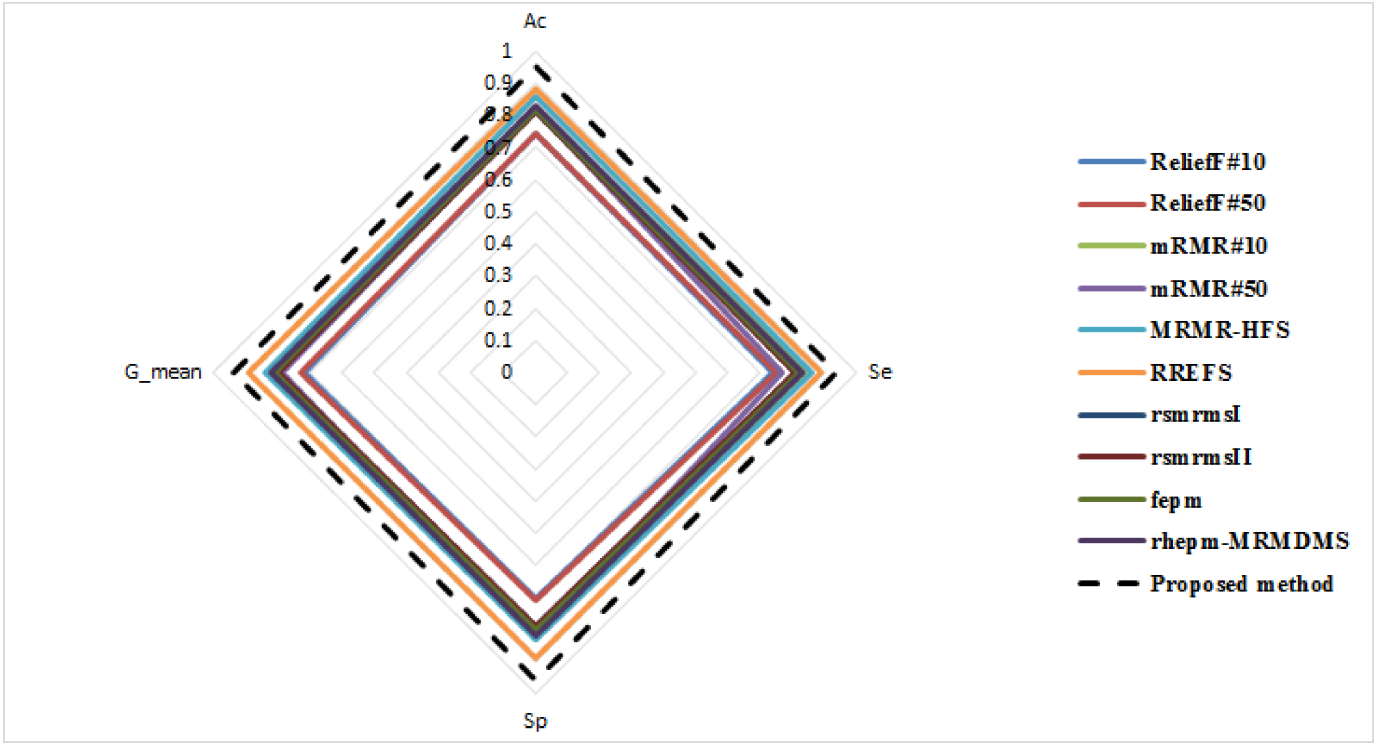

Figure 7: Radar chart of Naive Bayes classifier results based on evaluation measures.

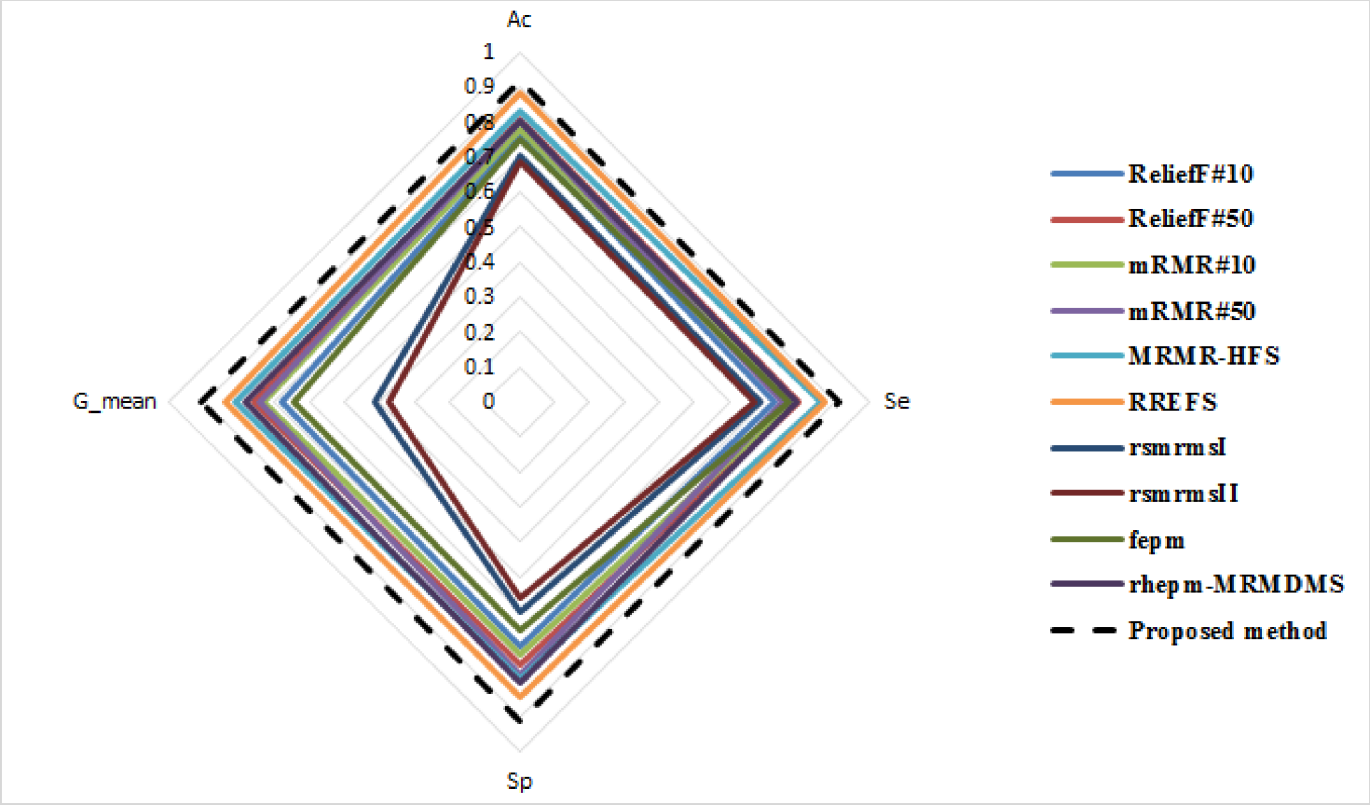

Figure 8: Radar chart of SVM classifier results based on evaluation measures.

- Both the common rough set and fuzzy-rough set concepts were utilized in the proposed method. The dependency degree for discretized features as well as its fuzzy version for continuous features were employed for feature selection.

There are several hesitant aggregating operators that can be used instead of information energy for fusing feature selection metrics in the future. Different rough based feature selection measures can be seen as some experts that have various opinions about a subset of features and these opinions can be fused via an aggregation operator of hesitant fuzzy sets. As another future work, one can combine different rough based feature selection metrics with the common filter based measures like the ReliefF and etc. Furthermore, common wrapper approaches and rough based wrapper methods can be combined via hesitant fuzzy aggregators for feature selection.

References

[1] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, *A review of feature selection methods on synthetic data*, Knowledge and information systems, **34**(3) (2013), 483–519.

[2] V. Bolón-Canedo, N. Sánchez-Marono, A. Alonso-Betanzos, J. M. Benítez, F. Herrera, *A review of microarray datasets and applied feature selection methods*, Information Sciences, **282** (2014), 111–135.

[3] G. Chandrashekar, F. Sahin, *A survey on feature selection methods*, Computers & Electrical Engineering, **40**(1) (2014), 16–28.

[4] N. Chen, Z. Xu, M. Xia, *Correlation coefficients of hesitant fuzzy sets and their applications to clustering analysis*, Applied Mathematical Modelling, **37**(4) (2013), 2197–2211.

[5] Y. Chen, Y. Xue, Y. Ma, F. Y, *Measures of uncertainty for neighborhood rough sets*, Knowledge-Based Systems, **120** (2017), 226–235.

[6] Y. Chen, Z. Zhang, J. Zheng, Y. Ma, Y. Xue, *Gene selection for tumor classification using neighborhood rough sets and entropy measures*, Journal of Biomedical Informatics, **67** (2017), 59–68.

[7] B. Choi, H. Kim, W. Cha, *A Comparative Study on Discretization Algorithms for Data Mining*, Communications for Statistical Applications and Methods, **18**(1) (2011), 89–102.

[8] A. Chouchoulas, Q. Shen, *Rough set-aided keyword reduction for text categorization*, Applied Artificial Intelligence, **15**(9) (2001), 843–873.

[9] J. Dai, Q. Xu, *Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification*, Applied Soft Computing, **13**(1) (2013), 211–221.

[10] M.K. Ebrahimpour, M. Eftekhari, *Ensemble of feature selection methods: A hesitant fuzzy sets approach*, Applied Soft Computing, **50** (2017), 300–312.

[11] M. K. Ebrahimpour, M. Zare, M. Eftekhari, G. Aghamolaei, *Occam's razor in dimension reduction: Using reduced row Echelon form for finding linear independent features in high dimensional microarray datasets*, Engineering Applications of Artificial Intelligence, **62** (2017), 214–221.

[12] U. M. Fayyad, K. B. Irani, *Multi-interval discretization of continuous-valued attributes for classification learning*, in: Proceedings of the International Joint Conference on Uncertainty in AI, Chambery, France, **6**(1) (1993), 1022–1029.

[13] I. Guyon, A. Elisseeff, *An introduction to variable and feature selection*, Journal of machine learning research, **3**(1) (2003), 1157–1182.

[14] M. A. Hall, *Correlation-based feature selection for machine learning*, University of Waikato Hamilton, (1999), 51–151.

[15] R. Jensen, Q. Shen, *New approaches to fuzzy-rough feature selection*, IEEE Transactions on Fuzzy Systems, **17**(4) (2009), 824–838.

[16] K. Kaneiwa, *A rough set approach to multiple dataset analysis*, Applied Soft Computing, **11**(2) (2011), 2538–2547.

[17] I. Kononenko, *Estimating attributes: analysis and extensions of RELIEF*, European conference on machine learning, (1994), 171–182.

[18] M. Kudo, J. Sklansky, *Comparison of algorithms that select features for pattern classifiers*, Pattern recognition, **33**(1) (2000), 25–41.

[19] J. Liu, Q. Hu, D. Yu, *A comparative study on rough set based class imbalance learning*, Knowledge-Based Systems, **21**(8) (2008), 753–763.

[20] J. Liu, Q. Hu, D. Yu, *A weighted rough set based method developed for class imbalance learning*, Information Sciences, **178**(4) (2008), 1235–1256.

[21] P. Maji, *A Rough Hypercuboid Approach for Feature Selection in Approximation Spaces*, IEEE Transactions on Knowledge and Data Engineering, **26**(1) (2014), 16–29.

[22] P. Maji, P. Garai, *On fuzzy-rough attribute selection: criteria of max-dependency, max-relevance, min-redundancy, and max-significance*, Applied Soft Computing, **13**(9) (2013), 3968–3980.

[23] P. Maji, S. K. Pal, *Fuzzy Rough Sets for Information Measures and Selection of Relevant Genes From Microarray Data*, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), **40**(3) (2010), 741–752.

[24] P. Maji, S. Paul, *Rough set based maximum relevance-maximum significance criterion and Gene selection from microarray data*, International Journal of Approximate Reasoning, **52**(3) (2011), 408–426.

[25] P.E. Meyer, *Information-theoretic variable selection and network inference from microarray data*, Ph. D. Thesis. Université Libre de Bruxelles, (2008), 19–84.

[26] M. Moradkhani, A. Amiri, M. Javaheri, H. Safari, *A hybrid algorithm for feature subset selection in high-dimensional datasets using FICA and IWSSr algorithm*, Applied Soft Computing, **25** (2015), 123–135.

[27] J. Moreno-Torres, J. Sáez, F. Herrera, *Study on the impact of partition-induced dataset shift on k-fold cross-validation*, IEEE Transactions on Neural Networks and Learning Systems, **23**(8) (2012), 1304–1312.

[28] Z. Pawlak, *Rough sets*, International Journal of Parallel Programming, **11**(5) (1982), 341–356.

[29] H. Peng, F. Long, C. Ding, *Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy*, IEEE Transactions on pattern analysis and machine intelligence, **27**(8) (2005), 1226–1238.

[30] D. Sheskin, *Handbook of parametric and nonparametric statistical procedures*, crc Press, (2003), 225–239.

[31] A. Statnikov, I. Tsamardinos, Y. Dosbayev, C. F. Aliferis, *GEMS: a system for automated cancer diagnosis and biomarker discovery from microarray gene expression data*, International journal of medical informatics, **74**(7) (2005), 491–503.

[32] V. Torra, *Hesitant fuzzy sets*, International Journal of Intelligent Systems, **25**(6) (2010), 529–539.

[33] E. Tuv, A. Borisov, G. Runger, K. Torkkola, *Feature selection with ensembles, artificial variables, and redundancy elimination*, Journal of Machine Learning Research, **10**(Jul) (2009), 1341–1366.

[34] H. Uğuz, *A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm*, Knowledge-Based Systems, **24**(7) (2011), 1024–1032.

[35] C. Wang, M. Shao, Q. He, Y. Qian, Y. Qi, *Feature subset selection based on fuzzy neighborhood rough sets*, Knowledge-Based Systems, **111** (2016), 173–179.

[36] X. Zhang, C. Mei, D. Chen, J. Li, *Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy*, Pattern Recognition, **56** (2016), 1–15.