

Available online at <http://www.meecspress.net/ijmsc>

An Enhanced Rough Set based Feature Grouping Approach for Supervised Feature Selection

Rubul Kumar Bania^a

^aAssistant Professor, Department of Computer Applications, North-Eastern Hill University, Tura Campus, Meghalaya, India

Received: 08 August 2017; Accepted: 16 October 2017; Published: 08 January 2018

Abstract

Selection of useful information from a large data collection is an important and challenging problem. Feature selection refers to the problem of selecting relevant features from a given dataset which produces the most predictive outcome as the original features maintain before the selection. Rough set theory (RST) and its extension are the most successful mathematical tools for feature selection from a given dataset. This paper starts with an outline of the fundamental concepts behind the rough set and fuzzy rough set based feature grouping techniques which are related to supervise feature selection. Supervised Quickreduct (QR) and fuzzy-rough feature grouping Quickreduct (FQR) algorithms are highlighted here. Then an enhanced version of FQR method is proposed here which is based on rough set dependency criteria with feature significance measure that select a minimal subset of features. Also, the termination condition of the base method is modified. Experimental studies of the algorithms are carried out on five public domain benchmark datasets available in UCI machine learning repository. JRip and J48 classifier are used to measure the classification accuracy. The performance of the proposed method is found to be satisfactory in comparison with other methods.

Index Terms: Feature selection, lower approximation, fuzzy set, rough set.

© 2018 Published by MECS Publisher. Selection and/or peer review under responsibility of the Research Association of Modern Education and Computer Science

1. Introduction

The value of knowledge can be recognized properly when it can be used efficiently and effectively. So, knowledge discovery is increasingly being acknowledged as a key element in extracting its value. The impact of data abundance is extending in recent years, in different areas and disciplines as varied as science, sports, e-commerce, advertising. There is a move toward data-driven knowledge discovery and decision-making. Data

* Corresponding author. Tel.: +91-9613424966

E-mail address: rubul.bania@gmail.com

reduction is an important step in knowledge discovery from large datasets [1][2]. The high dimensionality of dataset can be reduced by using suitable techniques, depending on the requirements of the data mining approach. The techniques fall into one of the two categories: those that transform the underlying meaning of the data features and those that are semantics-preserving [3][4]. Feature selection (FS) methods belong to the latter category, where a smaller set of the original features is chosen based on a subset evaluation criteria [2]. Rough set theory is an efficient mathematical technique for knowledge discovery from datasets [1]. This theory was initially developed for a finite universe of discourse in which the knowledge base is a partition, which is obtained by any equivalence relation defined on the universe of discourse. From a given a data set with feature values, it is possible to find a subset known as the *reduct* set, of the original features using RST that are the most informative. All other attributes can be removed from the data set with minimal information loss. From the dimensionality reduction perspective, informative features are those that are most useful in determining classifications from their values. The main advantage of rough set analysis is that it requires no additional parameters to operate other than the supplied data [4]. It works by making use of the granularity structure of the data only. This is a major difference when compared with Dempster-Shafer theory and fuzzy set theory which requires probability assignments and membership values, respectively [3][8]. The training data use for data mining applications and for learning approaches can be either labelled or unlabeled. It emerges the development of supervised, unsupervised FS algorithms. Supervised FS determines feature relevance by calculating feature's correlation or dependency with the given class labels, whereas unsupervised FS exploits data variation and separability to evaluate feature relevance without the class labels [2].

Feature selection using feature grouping technique is one of the new area which was first introduced in [3] and fuzzy set theory concept is applied for fuzzification of dependency values of attributes [4]. It was applied on fuzzy-rough set based Quickreduct algorithm. The main idea behind that approach is to speed up the reduct search process. In the literature there are some other associated works related to rough set feature selection. Velayutham et. al.[6] presented a method which uses dependency criterion with informed search methods. Liang et. al.[5] proposed an approach which can deal with large scale data by dividing data into multi sub-table granularity view. Given a large-scale dataset, the algorithm first selects different small granularities and then estimate on each small granularity the reduct of the original data set. In [7] authors have attempted to make theoretical comparative review on some existing RST methods with their pros and cons. Three RST based feature selection algorithms like Quickreduct (QR), Relative Reduct (RR) and Entropy based Reduct (EBR) were applied on six public domain datasets available in UCI machine learning repository.

The remaining parts of the paper are structured as follows. In Section 2 describes the theoretical background behind feature selection and basic ideas of RST. Section 3 describes the existing RST base Quickreduct feature selection algorithm. Section 4 introduces the Fuzzy-rough feature grouping ideas and algorithm. Section 5 explains the proposed algorithms with worked example. Sections 6 discuss the experimental results. Section 7 concludes the paper.

2. Theoretical Background

Rough set model is capable of dealing with feature selection process. To clarify the relationships among them, this section reviews some basic concepts, which facilitates the understanding of the remainder of this paper.

2.1. Feature Selection

Attribute selection (also known as subset selection) [3][4] is the process to determine a minimal feature subset from an input problem domain while retaining a suitably high accuracy in representing the original features. Selected subset contains the least number of dimensions that mostly contribute to accuracy; by discarding the remaining, unimportant dimensions. Let us consider a finite feature set $X = (x_1, x_2, x_3, \dots, x_M)$ of M -dimension of features with a given output class label $l_c \in L$, where $L = (l_1, l_2, \dots, l_C)$ is the set of the output class

labels. The minimal subset representation of X is say $= (x_1, x_2, x_3, \dots, x_N)$ where $N \leq M$ which maintain higher accuracy level of the same class label l_c as the original set X by optimizing a criterion function $f(Y)$. After getting the reduced set it is beneficial to use it in the real world mining problems which increase the efficiency and accuracy of resulting output.

2.2. Basics of Rough Set Theory

Rough set theory (RST) [1][9] is based on two important concepts, an upper and a lower approximation of a set. The lower approximation is a description of the domain objects which are known with certainty to belong to the subset of interest, whereas the upper approximation is a description of the objects which possibly belong to the subset.

2.2.1. Information System & Indiscernible Relation

Let $IS = \langle U, A, V, f \rangle$ be an information system (also known as information table) where U is the universe of discourse with a non-empty set of finite objects [13]. A is a non-empty finite set of attributes, which can be divided into two disjoint set conditional attribute set C and decision attribute set D , i.e., $A = \{C \cup D\}$ and $C \cap D = \emptyset$. V is the union of attribute domains, $V = \bigcup_{a \in A} V_a$, where V_a is the value set of attribute a , called the domain of a , $f: U \times A \rightarrow V$ is the total decision function called information function such that $f(x, a) \in V_a$, for all $a \in A, x \in U$.

An example of information system in Table 1 is shown, where $U = \{O1, O2, O3, O4, O5, O6, O7, O8, O9\}$, $C = \{a1, a2, a3, a4\}$, $D = \{d1\}$, $V = \{0,1\}$.

Table 1. An example Information System

$x \in U$	a1	a2	a3	a4	d1
O1	1	0	0	0	0
O2	1	1	0	0	1
O3	0	0	1	0	1
O4	0	1	0	1	0
O5	0	1	1	0	1
O6	1	0	1	1	0
O7	1	1	1	0	1
O8	1	1	0	0	1
O9	0	1	0	0	0

Information system [1] contains knowledge about the universe in terms of a predefined set of attributes. A subset of objects of the universe is called a concept in rough set theory. In order to represent or approximate these concepts, an information granule or equivalence relation is defined. In the context RST, it is termed as equivalence relation as an indiscernibility relation. If $P \subseteq A$, there is an associated equivalence relation:

$$IND(P) = \{(x, y) \in UXU : \forall a \in P, f(a, x) = f(a, y)\} \tag{1}$$

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

2.2.2. Lower and Upper Approximation

Let $X \subseteq U$ can be approximated using only the information contained within P by constructing the P -

lower approximation $\underline{P}X$ and P -upper approximation $\overline{P}X$ of set X can be defined as:

$$\underline{P}X = \bigcup_{x \in U} \{[x]_P \mid [x]_P \subseteq X\} \quad (2)$$

$$\overline{P}X = \bigcup_{x \in U} \{[x]_P \mid [x]_P \cap X \neq \emptyset\} \quad (3)$$

It is such a tuple $(\underline{P}X, \overline{P}X)$ that is called a rough set.

Let, $Q \in A$, be equivalence relations over U , then the Positive region can be defined as:

$$POSR_P(Q) = \underline{P}X \quad (4)$$

2.2.3. Dependency Degree and Significance Measure

An important issue in data analysis is discovering *Dependencies* between attributes. Various measures can be defined to represent how much Q , a set of decision attributes, depends on P , a set of condition attributes. One of the most common measure is the dependency [1] denoted by following way:

$$\gamma_P(Q) = \frac{|POSR_P(Q)|}{|U|} \quad (5)$$

If $\gamma_P(Q) = 1$, Q depends totally on P . If $0 < \gamma_P(Q) < 1$, Q depends partially on P and if $\gamma_P(Q) = 0$ then P does not depend on P . By calculating the change in dependency when an attribute is removed from the set of considered possible attributes, an estimate of the *Significance* of that attribute can be obtained. The higher the change in dependency, the more significant [8] the attribute is. If the significance is 0, then the attribute is dispensable without losing information. More formally, given P , Q and an attribute $a \in P$, the significance of attribute a upon Q is defined by:

$$\sigma_P(Q, a) = \gamma_P(Q) - \gamma_{P-\{a\}}(Q) \quad (6)$$

In RST feature selection is known as attribute reduction. The goal of attribute reduct is to remove redundant attributes so that the reduced set provides the same quality of classification as the original. For this, the concept of a reduct set was defined as a minimal subset R of the initial attribute set C such that for a given set of attributes D , $\gamma_R(D) = \gamma_C(D)$. The reduct set R is a minimal subset if $\gamma_{R-\{a\}}(Q) \neq \gamma_R(D)$, $\forall a \in R$.

3. Rough Set based Quickreduct Algorithm

The Quickreduct (QR) algorithm [8] is a dependency function-based reduction approach that attempts to generate reduct set using a hill climbing search approach without exhaustively generating all possible subsets. The reduction of attributes is achieved by comparing equivalence relations generated by sets of attributes. Attributes are removed so that the reduced set provides the same predictive capability of the decision feature as the original. This procedure starts off with an empty set and adds in turn at a time, those attributes that result in the greatest increase in the rough set dependency metric, until this produces its maximum possible value for the dataset. The pseudo code of the Quickreduct is given in Algorithm 1:

Algorithm 1: QuickReduct

Input: C, the set of all conditional attributes and D, the set of decision attributes.

Output: Reduct set R

```

Step1: R ← { }
Step 2: Do
Step3:   T ← R
Step4:   For Each x ∈ (C - R)
Step5:     If  $\Upsilon_{R \cup \{x\}}(D) > \Upsilon_T(D)$ 
Step6:       Where  $\Upsilon_R(D) = \text{card}(\text{POS}_R(D)) / \text{card}(U)$ 
Step7:         T ← R ∪ {x}
Step8:     Else R ← T
Step9:   Until  $\Upsilon_R(D) = \Upsilon_C(D)$ 
Step10: Return R

```

The advantage of this algorithm is that it is easy to understand and for medium complete dataset it works fine.

From the disadvantages point, dependency measure between the attributes sometimes leads the search to a non minimal path. Not convergent to all real data and has a poor termination condition.

4. Fuzzy Set and Rough Set-based Feature Grouping Algorithm

Fuzzy rough set-based feature grouping method is proposed in [3][4] for selecting features from some given benchmark data. Authors have also mentioned that same method can be applied for rough set feature selection approach, only the dependency function applied will be by crisp rough sets. In general the degree of dependency value of an attribute always lies in the range [0, 1], with 0 indicating no dependency and 1 indicating total dependency. Let us take one example to mention the utility of feature grouping: say, two subsets of conditional attributes in a dataset may have the dependency degrees like $\Upsilon_{\{a1, a5, a3\}}^{(D)} = 0.89$ and $\Upsilon_{\{a1, a2, a3\}}^{(D)} = 0.87$. If we apply the Quickreduct feature selection algorithm then {a1, a5, a3} will be the optimal subset as it has higher dependency value than other one. But it may not be the scenario if the dataset is real and it has noise. It may be the case that {a1, a2, a3} is the best subset rather than {a1, a5, a3}. By fuzzifying [5] the output values of the dependency values this type of problem is successfully handled.

4.1. Fuzzy Dependency and Scaling

Several fuzzy sets [8][14] are defined over the dependency range i.e. [0 to 1]. In Fig.1 an example of fuzzification on dependency values with three fuzzy sets small(S), medium (M) and large (L) is shown.

Now, a scaling operation on the dependency values will be performed. As soon as the reduct subset is evaluated, the highest and lowest dependencies $\Upsilon_{high}(D)$ and $\Upsilon_{low}(D)$ are computed using equation 5 to scale the dependency degree for subset P.

$$\Upsilon'_P(D) = \frac{\Upsilon_P(D) - \Upsilon_{low}(D)}{\Upsilon_{high}(D) - \Upsilon_{low}(D)} \quad (7)$$

By this scaling method, the attribute subset with the highest dependency value will have a scaled dependency of 1. The subset with the lowest will have a scaled dependency of 0. The fuzzy-rough Quickreduct algorithm (FQR) which employs scaling and fuzzy dependencies is available in [3].

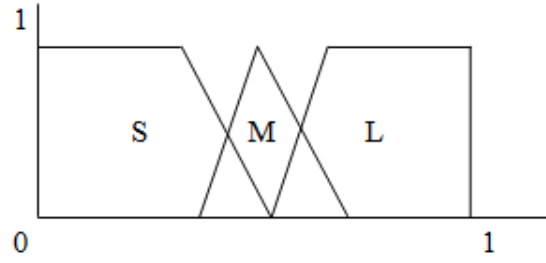


Fig.1. Fuzzification of Dependency Values

5. Proposed Rough Feature Grouping Algorithm

In this section a modified version of FQR is proposed which uses the rough set theory based dependency calculation. The use of ‘cands’ set is same as the FQR algorithm. In the feature selection process strong α -cut set [8] is used by setting α value for dependency value on the required label of fuzzy set, which selects attributes above it. After the attributes are selected the significance of the feature is calculated. This significance measure of attribute is not performed in FQR. If significance value is greater than 0, then the attribute is indispensable else dispensible and store those attribute in S1 which are indispensable. If the cardinality of S1 is less than S then reduct set R will hold S1 otherwise S. This process continues until the stopping criteria is fulfill, where P calculate a floating type value with the help of the dependency value of the current reduct set and the cardinality of the conditional attribute. This value is compare with a user input threshold value and it helps to reduce the search space. The threshold value in this work is set to 1. These changes are performed on the existing FQR algorithm and it is given in Algorithm 2. One sample of the fuzzification of dependency value with three fuzzy membership functions is shown in Fig 2.

$$\text{Small}(x) = \begin{cases} 0 & \text{when } x > 0.5 \\ (0.5 - x) / 15 & \text{when } 0.35 < x < 0.5 \\ 1 & \text{when } x < 0.35 \end{cases}$$

$$\text{Mid}(x) = \begin{cases} 0 & \text{when } x < 0.35 \text{ or } x > 0.65 \\ (0.5 - x) / 15 & \text{when } 0.35 < x < 0.5 \\ (0.65 - x) / 15 & \text{when } 0.5 < x < 0.65 \\ 1 & \text{when } x = 0.5 \end{cases}$$

$$\text{Large}(x) = \begin{cases} 0 & \text{when } x < 0.5 \text{ or } x > 1 \\ (0.65 - x) / 15 & \text{when } 0.5 < x < 0.65 \\ 1 & \text{when } x > 0.65 \end{cases}$$

Fig.2. Fuzzification of Dependency Values

Algorithm 2: Improve Fuzzy QuickReduct

Input: C, the set of all conditional attributes and D, the set of decision attributes.

Output: Reduct set $\bar{R}ed$

```

Step 1:  $\bar{R}ed \leftarrow \{\}$ 
Step 2: Do
Step 3:  $Y_{high} = 0; Y_{low} = 1$ 
Step 4:  $Cands \leftarrow \{\}$ 
Step 5: For Each  $x \in (C - \bar{R}ed)$ 
Step 6:  $T \leftarrow R \cup \{x\}$ 
Step 7:  $Cands \leftarrow Cands \cup (x, Y_T(D))$ 
Step 8: If  $Y_T(D) > Y_{high}$ 
Step 9:  $Y_{high} = Y_T(D)$ 
Step 10: else if  $Y_T(D) < Y_{low}$ 
Step 11:  $Y_{low} = Y_T(D)$ 
Step 12:  $Cands \leftarrow Scale(Cands, Y_{high}, Y_{low})$ 
Step 13:  $S \leftarrow selectfeatures(Cands)$ 
Step 14: For each  $x \in S$ 
Step 15: If  $\sigma_S(D, x) > 0$ 
Step 16:  $S1 \leftarrow S1 \cup \{x\}$ 
Step 17: If  $card(S1) < card(S)$ 
Step 18:  $\bar{R}ed \leftarrow \bar{R}ed \cup \{S1\}$ 
Step 19: Else  $\bar{R}ed \leftarrow \bar{R}ed \cup \{S\}$ 
Step 20:  $P = (1 - Y_R(D)) * card(s)$ 
Step 21: Until  $P > thres$ 
Step 22: For each  $x \in \bar{R}ed$ 
Step 23: If  $\sigma_R(D, x) > 0$ 
Step 24:  $\bar{R}ed \leftarrow x$ 
Step 25: Return  $\bar{R}ed$ 

```

Worked Example:

We are considering the artificial crisp dataset from Table 1 [6] to explain the execution steps of the proposed algorithm. The α value used in feature selection method is set to 0.5. For each attribute dependency is calculated and store in cans which is shown Table 2. After that the scaling of the attributes are calculated with the help of equation 7.

$$Y'_p(D) = Y_p(D) - Y_{low}(D) / Y_{high}(D) - Y_{low}(D), \text{ For attribute } a4,$$

$$\text{Here } Y_p(D) = 0.2222, Y_{low}(D) = 0.0 \text{ and } Y_{high}(D) = 0.2222$$

$$Y'_p(D) = 0.2222 - 0.0000 / 0.2222 - 0.0000 = 1$$

Similarly for other attributes a1, a2 and a3 scaling is done and all of them has scaling value equal to 0. So, all of them are labelled on the defined fuzzy sets with corresponding scaling values which is shown in Table 3. The feature selection procedure will return subset {a4} as it is in the Large label in store in S.

Now for each value in S, the significance of the attribute will be calculated as below.

$$\sigma_T(D, x) = Y_T(D) - Y_{T-\{x\}}(D), \text{ Here, } T = \{d\}, D = \{d\}, x = \{a4\}$$

$$\sigma_T(D, a4) = Y_{\{a4\}}(\{d\}) - Y_{\{d\}}(d) = 0.2222 - 0.000 = 0.2222 > 0$$

As the significance is greater than 0, attribute d is not dispensable. The value of P in this state will be calculated as $P = (1 - 0.2222) * 4 = 3.1111 > 1$, so execution continue.

Table 2. Dependency Values

Subset	{a1}	{a2}	{a3}	{a4}
Dependency	0	0	0	0.222

Table 3. Scaling with Membership Value

Subset	{a1}	{a2}	{a3}	{a4}
Scaling	0.000	0.000	0.000	1.000
Fuzzy Label (membership)	Small = 1.0 Mid = 0.0 Large = 0.0	Small = 1.0 Mid = 0.0 Large = 0.0	Small = 1.0 Mid = 0.0 Large = 0.0	Small = 0.0 Mid = 0.0 Large = 1.0

In the next iteration for subset {a1, a4}, {a2, a4} and {a3, a4} dependency and for each attribute scaling will be calculated using the same way as above. Below in Table 4 and Table 5 the result is shown.

Table 4. Dependency Values

Subset	{a1,a4}	{a2,a4}	{a3,a4}
Dependency	0.222	0.222	0.556

The feature selection procedure will return subset {a3} as it is in the Large label in store in S. Now for each attribute, a3 we have to calculate significance value.

$$\sigma_T(D, a3) = Y_{\{a3\}}\{d\} - Y_{\{ \}}(d) = 0.0000 - 0.0000 = 0$$

a3 significant value which is not greater than 0, so S1 set will not store the attributes. Now, as the cardinality of S1 is not less than S, the reduct set will store S. The reduct Subset will be {a3, a4}. The value of P in this state will be calculated as $P = (1 - 0.5556) * 4 = 1.77778 > 1$, so execution continue.

Table 5. Scaling with Membership Value

Subset	{a1}	{a2}	{a3}
Scaling	0.000	0.000	1.000
Fuzzy Label (membership)	Small = 1.0 Mid = 0.0 Large = 0.0	Small = 1.0 Mid = 0.0 Large = 0.0	Small = 0.0 Mid = 0.0 Large = 1.0

In the next iteration for subset {a1, a3, a4}, {a2, a3, a4} dependency and for each attribute scaling will be calculated using the same way as above. Table 6 and Table 7 show the result.

Table 6. Dependency Values

Subset	{a1,a3, a4}	{a2,a3,a4}
Dependency	0.6667	0.6667

Table 7. Scaling with Membership Value

Subset	{a1}	{a2}
Scaling	0.000	0.000
Fuzzy Label (membership)	Small = 0.0 Mid = 0.0 Large = 1.0	Small = 0.0 Mid = 0.0 Large = 1.0

The feature selection procedure will return subset {a1, a2} as it is in the Large label in store in S. Now for each attribute, a1, a2 we have to calculate significance value.

$$\sigma_T(D, a1) = Y_{\{a1,a2\}}\{d\} - Y_{\{a1\}}(d) = 0.6667 - 0.0000 = 0.6667 > 0$$

$$\sigma_T(D, a2) = Y_{\{a1,a2\}}\{d\} - Y_{\{a2\}}(d) = 0.6667 - 0.0000 = 0.6667 > 0$$

{a1, a2} significant value which greater than 0, so S1 set will store the subset. Now, as the cardinality of S1 is not less than S, the reduct set it will store S. The reduct Subset will be {a1, a2, a3, a4}. The value of P in this state will be calculated as $P = (1 - 1) * 4 = 0 < thres$, so loop will be terminated.

Now, the reduct subset is {a1, a2, a3, a4} but again for each attribute a1, a2, a3 and a4 the significant value will be.

$$\sigma_T(D, a1) = Y_{\{a1,a2,a3,a4\}}\{d\} - Y_{\{a2,a3,a4\}}(d) = 1.0000 - 0.6667 = 0.3333 > 0$$

$$\sigma_T(D, a2) = Y_{\{a1,a2,a3,a4\}}\{d\} - Y_{\{a1,a3,a4\}}(d) = 1.0000 - 0.6667 = 0.3333 > 0$$

$$\sigma_T(D, a3) = Y_{\{a1,a2,a3,a4\}}\{d\} - Y_{\{a1,a2,a4\}}(d) = 1.0000 - 0.7778 = 0.2222 > 0$$

$$\sigma_T(D, a4) = Y_{\{a1,a2,a3,a4\}}\{d\} - Y_{\{a1,a2,a3\}}(d) = 1.0000 - 1.0000 = 0$$

Since a4 has significance 0 it is dispensable. But attributes a1, a2 and a3 are indispensable. So the final reduct set will be {a1, a2, a3}.

6. Experimental Results

This section presents the results of the experimental work which are performed on the existing and proposed method. The benchmark datasets are taken from the UCI Repository of Machine Learning repository [15]. In Table 8 details of the five (5) datasets are shown. column1, column2, column3 and column4 represents the dataset name, number of objects it contains, number of features and number of classes respectively. The classifier tool WEKA [10] is used for classification process. WEKA is an open source java based machine-learning workbench. It brings together many machine learning algorithms for classification process under a common umbrella. After the feature reduced set generated by the above mentioned algorithms, two classifier learners JRip and J48 are employed for the classification task. J48 is open source implementation of the C4.5 algorithm and JRip is a rule based classifier [10]. The average classification accuracy of individual classifier in terms of percentage is obtained by using 10-fold cross validation [11]. The classification was initially performed on the unreduced dataset, followed by the reduced datasets.

Table 8. Details of Data set

Dataset	Objects	Features	Class
Iris	150	4	3
Wine	178	13	3
Lung cancer	32	56	2
Heart	270	14	2
Pima	768	8	2

6.1. Comparison of Performance between Supervised QR, MQR and FQR

From the unreduced datasets, features or attributes are selected by the discussed supervised QR, rough set based FQR and proposed IFQR algorithm. The total numbers of features and individual feature numbers which are selected by these algorithms are tabulated in Table-9.

The performance measures used for the evaluation of the feature selection methods in terms of chosen classifier include accuracy, recall (sensitivity) and specificity [12]. These measures are defined in terms of confusion matrix elements TP (true positive), TN (true negative), FP (false positive), and FN (false negative) as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (8)$$

$$Recall(Sensitivity) = \frac{TP}{TP+FN} \quad (9)$$

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

The unreduced datasets were first applied to the classifiers, JRIP and J48. Then the reduced data sets which were obtained by using the QR, FQR and IFQR methods were applied for finding the classification accuracy. All the results are presented in terms of classification accuracy and it is tabulated in Table-10. Also the performance measure of classification by recall (sensitivity) and precision is calculated.

For Iris dataset QR and FQR has selected same numbers of attributes, whereas IFQR has selected same as unreduced dataset. Classification accuracy has degraded by QR and FQR for both the classifier. It is interesting to note that there is a slight increase in classification accuracy for the QR, FQR and IFQR, with respect to the unreduced Pima dataset, for J48 classifier but for JRip it is decreased. For wine dataset IFQR performance in terms of accuracy is better than the others. In case of Lung cancer dataset FQR, IFQR accuracies got slightly increased with comparison to QR methods. For Heart dataset QR, FQR and IFQR performance is poor in comparison to original unreduced data. In below Table 11 shows the calculations of some standard measures such as precision and recall for JRip classifier. Similarly in Table 12 the calculations of some standard measures such as precision and recall for J48 classifier are shown.

Table 9. Comparison of selected Features by Three Different Methods

Dataset	QR	FQR	IFQR
Iris	3[1,2,3]	3[1,2,3]	4[1,2,3,4]
Wine	5[1,7,8,11,12]	5[1,7,9,11,12]	4[1,7,9,12]
Lung cancer	9[2,3,5,6,12,13,23,40,41]	9[2,3,5,6,12,13,23,40,43]	8[2,3,5,6,12,13,40,41]
Heart	3[1,4,5]	5[1,4,5,6,11]	5[1,4,5,6,11]
Pima	4[1,2,,6,7]	4[1,2,,6,7]	4[1,2,6,7]

Table 10. Comparison of Classification Accuracy by Different Methods

Algorithm	Classifier	Iris	Wine	L.Cancer	Heart	Pima
Unreduced Data	JRip	95.33	92.17	78.12	78.88	76.04
	J48	96.00	91.53	78.04	76.67	73.82
QR	JRip	92.67	90.52	78.12	58.51	74.21
	J48	93.33	89.29	75.00	58.51	74.86
FQR	JRip	92.67	90.60	78.18	65.55	74.21
	J48	93.33	89.83	78.12	67.41	74.86
IFQR	JRip	95.33	92.52	78.28	65.55	74.21
	J48	96.00	91.87	78.34	67.41	74.86

Table 11. Calculated Measure of Recall, Precision by JRip Classifier

Algorithm	Measure	Iris	Wine	L.Cancer	Heart	Pima
QR	Recall	0.980	0.743	0.751	0.647	0.842
	Precision	1.000	0.687	0.676	0.622	0.780
FQR	Recall	0.980	0.781	0.741	0.693	0.842
	Precision	1.000	0.722	0.626	0.689	0.780
IFQR	Recall	1.000	0.744	0.715	0.693	0.842
	Precision	1.000	0.721	0.708	0.689	0.780

Table 12. Calculated Measure of Recall, Precision by J48 Classifier

Algorithm	Measure	Iris	Wine	L.Cancer	Heart	Pima
QR	Recall	0.980	0.743	0.751	0.593	0.832
	Precision	1.000	0.687	0.656	0.636	0.792
FQR	Recall	0.980	0.781	0.731	0.696	0.832
	Precision	1.000	0.722	0.616	0.733	0.790
IFQR	Recall	1.000	0.744	0.775	0.696	0.832
	Precision	1.000	0.721	0.711	0.733	0.790

7. Conclusion & Future Work

Feature selection for large data sets is still a challenging issue in the field of pattern recognition, data mining. In this work, rough set based feature selection with feature grouping techniques is discussed. The modified version of rough set feature grouping Quickreduct with fuzzy set is proposed. Based on the experiment result the output of classification accuracy for a few input dataset is quite satisfactory. In future the performance of this proposed method will be computed with some other real world datasets.

References

- [1] Z. Pawlak. Rough Sets, International J. Computer and Information Sciences, Vol. 11, No.5, pp.341-356, 1982.

- [2] R. Bello and J. L. Verdegay, Rough sets in the Soft Computing environment. Elsevier Information sciences, Vol. 212, pp.1-14, 2012.
- [3] R.W. Swiniarski and A. Skowron. Rough Set Methods in Feature Selection and Recognition, Pattern Recognition Letters, Vol. 24, No. 6, pp. 833-849, 2003.
- [4] Richard Jensen and Qiang Shen. "Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough-Based Approaches", IEEE transactions on knowledge and data engineering, Vol.16, No.12, pp.1457-1471, 2004.
- [5] R .Jensen and Q .Shen. Using Fuzzy Dependency - Guided Attribute Grouping in Feature Selection, Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Proc. Ninth Int'l Conf. pp.250-255, 2003.
- [6] L. Liang. An efficient rough feature selection algorithm with a multi-granulation view. International Journal of Approximate Reasoning, Vol.53, pp.912–926, 2012.
- [7] C. Velayutham and K. Thangavel. Improved Rough set algorithms for optimal attribute reduct. Journal of Electronic Science and Technology, Vol. 9, No.2, pp.108-117, 2011.
- [8] R.K.Bania, Comparative Review on Classical Rough Set Theory based Feature Selection Methods. International Journal of Computer Applications. Vol.114, No.19, pp.31-35. 2015.
- [9] D. Dubois and H. Prade, Rough fuzzy sets and fuzzy rough sets. International Journal of General Systems. Vol-17, No.2, pp.191-209, 1990.
- [10] N. M. Parthalaïn and Q .Shen, On rough sets, their recent extensions and applications. The Knowledge Engineering Review, Cambridge University Press, Vol. 25, No.4, pp.365–395, 2010.
- [11] WEKA <http://www.cs.waikato.ac.nz/ml/weka>.
- [12] J.Han and M.Kamber. Data Mining Concepts and Techniques 3rd Edition Morgan Kaufmann Publishers. 2012
- [13] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. Information Processing and Management, Elsevier. Vol.42. pp. 427–437, 2009.
- [14] E. Sallam, T. Medhat, A.Ghanem and M. E. Ali, Handling Numerical Missing Values via Rough Sets. International journal of Mathematical Sciences and Computing, Vol.2, pp.22-36, 2017.
- [15] D.B.Patil and Y.V. Dongre, A Fuzzy Approach for Text Mining. International journal of Mathematical Sciences and Computing, Vol.4, pp.34-43, 2015.
- [16] C.L.Blake and C.J.Merz. UCI Repository of machine learning databases.[Online].Available: <http://www.ics.uci.edu/~mllearn/>

Authors' Profiles



Rubul Kumar Bania, is an assistant professor in the Department of Computer applications, North-Eastern hill university, Tura campus, Meghalaya. He did his MCA and M.Tech-IT from Tezpur University, Assam. His research area of interest is Data mining, Image processing and Artificial intelligence.

How to cite this paper: Rubul Kumar Bania,"An Enhanced Rough Set based Feature Grouping Approach for Supervised Feature Selection", International Journal of Mathematical Sciences and Computing(IJMISC), Vol.4, No.1, pp.71-82, 2018.DOI: 10.5815/ijmsc.2018.01.05