ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2022.0391

ENTROPY BASED GREEDY UNSUPERVISED FEATURE SELECTION METHOD USING ROUGH SET THEORY FOR CLASSIFICATION

Rubul Kumar Bania[1] and Satyajit Sarmah[2]
[1]Department of Computer Application, North-Eastern Hill University, India
[2]Department of Information Technology, Gauhati University, India

Abstract

Feature selection technique attempts to select and remove irrelevant features while ensuring that an informative subset of features remains in the dataset. The performance of a classifier often depends on the feature subset used for the robust classification task. In the medical and healthcare application domain, classification accuracy plays a vital role. The higher level of false negatives in medical diagnosis systems may raise the risk of patients not employing the necessary treatment they need. In this article, we have proposed an unsupervised feature selection method that underlines the concepts of rough set theory for the task of classification of high-dimensional datasets. Experiments are carried out on seven public domain healthcare and life science related datasets. The obtained experimental results justify the significance of the proposed method over five other state-of-the-art feature selection methods.

Keywords:

Feature Selection, Rough Set, Unsupervised, Entropy

1. INTRODUCTION

The enrichment of technologies has amplified the size of the storage capabilities, and advances in data collection have led to loads of information. Medical/ healthcare data contains many different values regarding objects/patterns and features/attributes in many applications, such as gene expression data. These enormous data volumes far outpace human capability to comprehend and are challenging to handle. Moreover, one of the fundamental problems with these high dimensional datasets is that all the features may not be so informative, and it varies with applications.

Data reduction (DR) [1, 2, 3] is one of the data prepossessing techniques in anticipation of discovering the knowledge from the extensive data collection. In the literature, DR problem is handled using two approaches: transformation-based and selection-based [3]. The first approach, encode the basic meaning of the features, and is known as feature extraction [4] and the second approach maintains to holds the meanings of the actual content without transforming the underline meaning of the features, and it is known as feature selection (FS) [2, 3] approach. FS algorithms are often applied in data mining, pattern recognition, and machine learning problems to reduce the dimensionality of a dataset, which often describes the sample using thousands of features. The goal is to minimize the trained models' complexity, increase their accuracy, and reduce their consumption of computing resources [3, 4]. In the process of FS, a subset of the original feature set is selected based on a suitable subset evaluation criterion. The principal part of FS is to find out a minimal relevant feature subset from a given problem domain by obtaining high accuracy for representing the original features. It is possible, therefore, to broadly define each approach to FS based on some intrinsic characteristics such as the search strategy, generation of subsets, and feature subset evaluation measure. Moreover, in the literature, FS process is classified into four different categories viz., filter, wrapper, hybrid and embedded. The details of these techniques are broadly discussed in [3, 6, 7, 8, 13, 16].

However, based on the availability of the labels of the data, FS methods are usually categorized into two groups: supervised and unsupervised methods [3, 5, 6]. In the supervised FS (SFS) methods, the selection of the features is performed based upon the co-relation/dependency of the features with respect to the class label. On the contrary, in many of the high dimensional medical and healthcare applications data there is a high chance of scarcity of the class labels. This sort of issues may usually frustrate and increased the workloads of the domain experts to assign the labels for the patterns. However, it is also notable that not all the features are relevant or significant in order to assign a particular given label for the patterns. Therefore, for selecting the relevant features without considering the class labels lead towards the developments of unsupervised FS (UFS) methods. However, for the unsupervised FS task, as the decision class labels are not provided, it raises the question of *retaining of which features.* Moreover, not all the features of a dataset are very important and as some of them might be irrelevant, redundant, or noisy.

In the literature, rough set theory (RST) [9, 10, 11] is found out to be as a successful approximation-based mathematical model to knob the limitations in knowledge, i.e., imprecision and uncertainty. RST is a highly demanding approach for solving the problem of FS by employing the filter-based FS approach. It can assist in identifying and selecting the most highly significant features in a dataset. In the literature, different UFS methods with varying selection criteria have been proposed for the past few years. Some of the popular UFS methods, namely Laplacian score feature selection (LSFS) [7], multi-cluster feature selection (MCFS) [6] etc., are used in different applications. Moreover, in the literature by employing RST, very few works are available on the development of UFS models [14, 15, 19]. Additionally, RST in the information-theoretic framework is one of the recent advancements.

The contributions of this work are as follows:

- Proposed an unsupervised FS method which is underlined in the information theory framework of RST.
- Illustrative example of the proposed method is provided.
- On several real-life medical and healthcare datasets the experiments are carried out.

The remaining of the article is organized as follows. In Section 2 summarization of the related work and the theoretical background ideas of rough sets in information theoretic framework is given. In Section 3, demonstration of the proposed method with a description of the algorithm, as well as an illustrative example is given. Section 4 demonstrates the experimental results and discussions. Finally, concluding remarks and future works are highlighted in Section 5.

2. BACKGROUND STUDY

This section discusses a thorough overview of the literature survey and the essential concepts related to the proposed method.

2.1 RELATED WORK

The FS technique is very useful in the field of medical and healthcare data analysis [4, 12, 13, 20, 22]. In areas like prognosis, diagnosis, screening, etc., the decision-making processes could be implemented using machine-learning-based classification techniques. Thus, in medical applications, classification accuracy is very important. If the percentage of false negatives in screening systems is high, then there is a high chance that the patients will not get the attention they may require. Also, the higher the percentage of false alarms, the greater the worry and load on the medical resources. Several FS methods have been proposed in recent years by various researchers that work differently by applying various techniques (e.g., probability distribution, entropy, correlation, etc.) in the domain of medical and healthcare to different business analysis models. Inbarani et al. [12] have proposed a RST hybridization model based on the Particle Swarm Optimization with Relative Reduction (PSO-RR) and PSO-based Quick Reduction (PSO-QR) for the diagnosis of diseases like erythemato-squamous and breast cancer. Some of the other recent studies in FS techniques are summarized in Table.1, focusing on the medical/healthcare and life science datasets using various FS methods (supervised and unsupervised) and classification techniques.

Table.1. Summarized information of the state-of-the-art work

Contributors	Purpose	Techniques
Shilaskar et al. [17]	Feature selection and classification	The method initiates to locate a subset of feature by calculating the ranked of the features with a suitable distance measure then a forward selection, and backward elimination search techniques are applied to diagnosis cardiovascular disease using support vector machine (SVM) classifier.
Banu et al. [18]	Feature selection and classification	Supervised quick reduct (SQR), Entropy based reduct (EBR), and tolerance rough set based unsupervised feature selection methods are applied for Egyptian neonatal jaundice dataset. The performance of the method is evaluated on decision tree classifier.
Velayutham et al. [14]	Unsupervised Feature selection and classification	Proposed an unsupervised advanced version of the relative reduct algorithm using the RST dependency measure.
Velayutham et al. [15]	Unsupervised Feature selection and classification	On several benchmark data, unsupervised quick reduct (USQR) algorithm using RST is proposed. The quality of the reduced feature sets is evaluated in WEKA tool.
Yildirim et al. [27]	Feature selection and classification	This work mainly focuses on the effect of four FS methods namely information gain, ReliefF, One-R, and Principal component analysis for hepatitis data analysis using four different classifiers.
Jothi et al. [19]	Unsupervised Feature selection and classification	For the FS task, in this work authors have applied the USQR Hybrid Soft set based unsupervised Quick Reduct (SSUSQR).
Wang et al. [21]	Feature selection and classification	An algorithm called feature forest based on RST positive region is used to generate the feature reduct set on four medical datasets. Naïve bayes and SVM classifiers are used to evaluate the performance.
Nahato et al. [26]	Feature selection and classification	RST lower approximation concept with amalgamation of neural network is applied on several clinical dataset to predict the presence and absence of disease.

2.2 ROUGH SET THEORY IN INFORMATION-THEORETIC FRAMEWORK

Rough set theory (RST) [10, 11, 22] was proposed by Z. Pawlak in 1982 to strictly deals with uncertainty and incompleteness. The approximation space of a set such as upper and lower is the foundation of RST. In the recent literatures, the interpretation of the information theory concepts is also addressed in RST [9, 10]. Researchers have proved that the algebraic method of RST is like the RST under the information-theoretic framework [11].

Let us consider an information system $IS=<U,C,V,F>$ where $U=\{x_1,x_2,x_3\ldots x_n\}$ is the universe of discourse which have finite number of objects. C represents the finite set of features. For any $B\subseteq C$, the associated equivalence relation IND(B):

$$IND(B)=\{(x,y)\in U\times U\colon \forall a\in B, F(x,a) = F(y,a)\}$$

The partition of U generated by $IND(B)$ can be denoted by $U/IND(B)$. Now, let $X\subseteq U$, by generating the B-lower approximation $\underline{B}X$ and B-upper approximation $\overline{B}X$ of the concept set X, it can be approximated. Thus, it can be defined as:

$$\underline{B}X = \bigcup_{x\in U}\left\{[x]_B\,[x]_B \subseteq X\right\}$$

$$\overline{B}X = \bigcup_{x\in U}\left\{[x]_B\,[x]_B \cap X \neq \varnothing\right\}$$

The lower approximation is sometimes known as positive region, is the descriptions of the objects which are known with certainly belong to the concept of interest X, whereas the upper approximation is a description of the objects which possibly belong to X. It is such a tuple ($\underline{B}X$, $\overline{B}X$) is called as a rough set of X which is the representation of the concept set X_{in} the approximation space *IS* with respect to the equivalence relation. In Fig.1 approximation of set X is shown.

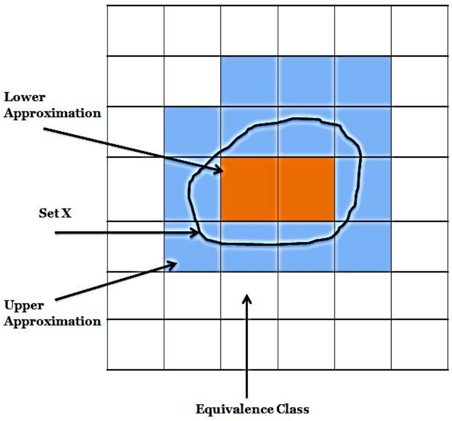


Fig.1. The approximations and regions of set X

If the partition of the universe is X, which is induced by P, where $X=U/IND(P) = \{X_1,X_2,X_3\ldots X_n\}$.

Then the distributions of the probability of X can be defined as:

$$[\boldsymbol{X};P]\begin{bmatrix} X_1 & X_2 & \ldots & X_n \\ P(X_1) & P(X_2) & \ldots & P(X_n) \end{bmatrix}$$

where $(X_i) = |X_i|/|U|$, $i = 1,2,\ldots n$. **|.|** denotes the cardinality of a set.

Definition 1: Given an information system *IS* and $P \subseteq C$ then $U/P=\{X_1,X_2,\ldots X_n\}$ is the condition partition. Then Shannon's entropy $H(P)$ of P can be defined as:

$$H(P) = -\sum_{i=1}^{n} P(X_i)\log P(X_i) = -\sum_{i=1}^{n} \frac{|X_i|}{|U|}\log\frac{|X_i|}{|U|}$$

Definition 2: If $P \subseteq C$ and $B \subseteq C$, $U/P=\{X_1,X_2,\ldots X_n\}$ and $U/B=\{Y_1,Y_2,\ldots Y_m\}$. Then conditional entropy of P conditioned to B can be defined as:

$$H(P|B) = -\sum_{i=1}^{n} P(X_i)\sum_{i=1}^{n} P(Y_j|X_i)\log_2 P(Y_j|X_i)$$

where $P(Y_j|X_i) = \frac{|X_i \cap Y_j|}{|X_i|}$. Then the conditional entropy can be defined as:

$$H(P|B) = -\sum_{i=1}^{n} \frac{|X_i|}{|U|}\sum_{j=1}^{m} \frac{|X_i \cap Y_j|}{|X_i|}\log_2 \frac{|X_i \cap Y_j|}{|X_i|}$$

3. PROPOSED METHODOLOGY

In this section the detailed description of the proposed methodology of this research is given and the block diagram is shown in Fig.2. Each step involved in this investigation are discussed below.

Step 1: The real-world data tend to be incomplete due to some technical issues, i.e., objects may not have any recorded values for some features. A significant presence of missing values in data may sometimes decrease the statistical power of the methods, and eventually, it reflects on the reliability of the results. In the literature, different approaches are adopted for handling such issues [22] - [25]. Observing some of the advantages, in this investigation, the weighted *k*-nearest neighbor (*k*-NN) imputation method is used to impute the missing values in the respective datasets. After imputing the missing values, those datasets with continuous (real valued) features are discretized [28] [29] to improve the knowledge comprehensibility. As, RST is more robust for discrete feature domains only, hence wherever it is applicable binning with equal width technique [29] is applied to the discretized continuous feature values before the FS process.

Step 2: After the preprocessing step, without including the class labels the dataset is passed to the proposed rough set with entropy for unsupervised feature selection (RE-UFS) method for generating a robust reduct set for an input dataset. Then the subset is passed to the classifiers for evaluating the predictive models. The detailed description of the method is given in the subsequent Subsection.

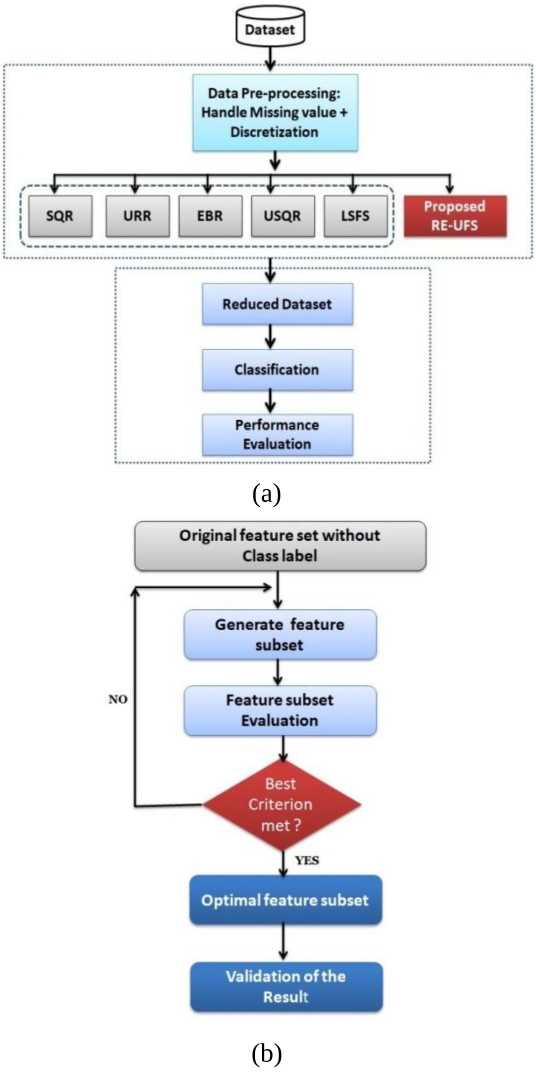
Step 3: To evaluate how the proposed method is performing, it is compared with the five state-of-the-art methods, namely SQR [11], EBR [15], URR [14], USQR [15], and LSFS [7]. First four methods (viz., SQR, EBR, URR and USQR) are based on the concept of RST. On the other hand, LSFS, is the Laplacian Score (LS) which is fundamentally underlined on concepts of laplacian eigen maps and locality preserving projection techniques. By calculating the locality preserving power LS evaluates the features. The reduct sets generated by the RE-UFS and compared state-of-the-art methods are then provided as an input to two popular state-of-the-art classifiers namely Support vector machine (SVM) [30] and Random Forest (RF) [31].

Step 4: Finally, the performance of the two classifiers is statistically checked using the average classification accuracy, precision, recall, F-measure, and Matthew's correlation coefficient measures**.** Then finally, the experimental results are compared to analyze and verify how the proposed RE-UFS method performed in comparison to the other five FS methods.

3.1 PROPOSED METHOD

Rough set with entropy for unsupervised feature selection (RE-UFS) algorithm iterates to find an informative feature subset without the generation of the exhaustive possible feature subsets. In Fig.2, schematic view of the proposed RE-UFS method is shown. The method begins with an empty set *RD*. Then it calculates the average entropy for each feature of a given dataset. It is worthy to mention that by applying the *Definition-1* and *Definition*-2 entropy of the features are calculated. Then it inserts a feature into the set *RD* if it has low entropy. The average entropy (*Mean* function) values of each feature are calculated, and as the process is greedy, so, in each stage, it finds out the best candidate who has a lower mean entropy value. The searching process of the RE-UFS method to select the best informative feature subset converges when the entropy value of the selected feature subset is similar to the complete feature set of the dataset. Usually, for a consistent dataset, the entropy value reaches zero. The different steps of the proposed algorithm are given below.

```
**Algorithm: RE-UFS**
**Input:** An Information system without the decision feature.
**Output:** Feature reduct set $RD$
$RD \leftarrow \emptyset$
While $Mean(H(x|RD)), \forall x \in C \neq Mean(H(x|C)), \forall x \in C$
$Temp \leftarrow RD$
While Mean($H$($x$ | $RD$)), ∀$x$∈$C$ ≠ $Mean(H(x|C))$, ∀$x$∈$C$ $Temp \leftarrow RD$
    for each $f \in (C\text{-}RD)$
        If $Mean(H(x|\{RD \cup f\}))$ ,$\forall x \in C$ < $Mean(H(x|Temp)),\forall x \in C$
            Then $Temp \leftarrow RD \cup \{f\}$
            $RD \leftarrow Temp$
        End If
    End for
End While
Return RD
```



(a)

(b)

Fig.2. (a) Block diagram of the proposed methodology (b) Schematic view of the execution process of RE-UFS method

By adopting the searching criterion from the literature [10] [11] [15] in this proposed method hill climbing approach is applied. The algorithm design technique follows in the Algorithm (RE-UFS) is greedy in nature. For which, in each iteration or in each stage the RE-UFS method attempts to apply a greedy choice which assumed to be the best at that stage and this process continues until a termination criterion is satisfied. Thus, RE-UFS is a greedy hill climbing approach to attain an informative feature subset. To understand the working principle of the method, one suitable example is demonstrated.

3.1.1 Illustrative Example for the RE-UFS Method:

Let us consider, a sample dummy dataset without having any decision feature (class label), which is shown in Table.2, where, where, U = {1,2,3,4,5,6,7}, C = {a,b,c,d}, and the value of V = {5,10,20}.

Table.2. Sample dummy dataset.

U	***a***	***b***	***c***	***d***
1	10	5	20	10
2	10	5	20	5
3	10	20	5	5
4	10	20	20	10
5	20	10	5	5
6	20	10	10	5
7	20	10	20	10

The indiscernability of the features {a},{b},{c} and {d} are calculated and tabulated in Table.3, Table.4, Table.5 and Table.6 respectively. In the Table.3, a_1, and a_2, represents the equivalence classes. Similarly, {b_1, b_2, b_3}, {c_1, c_2, c_3}, and {d_1, d_2, d_3} in the Table.4, Table.5, and Table.6 represents the equivalence classes.

Table 3. Indiscernibility for feature {a}

Feature	a_1	a_2
a	{1,2,3,4}	{5,6,7}

Table 4. Indiscernibility for feature {b}

Feature	b_1	b_2	b_3
b	{1,2}	{5,6,7}	{3,4}

Table 5. Indiscernibility for feature {c}

Feature	c_1	c_2	c_3
c	{3,5}	{6}	{1,2,4,7}

Table 6. Indiscernibility for feature {d}

Feature	d_1	d_2
d	{2,3,5,6}	{1,4,7}

Similarly, after calculating the average entropy for the features {b}, {c} and {d} results which are achieved is a shown in Table.7.

Table.7. Average entropy values for the features {a}, {b}, {c} and {d}

Feature	**{*a*}**	**{*b*}**	**{*c*}**	**{*d*}**
Average Entropy	0.42869	0.41105	0.51750	0.83962

Then select the feature with the lowest average entropy value. Thus, feature {b} is selected.

In the next step it is necessary to calculate the average entropy of all the subsets containing the feature {b} and other features. The indiscernability of the subsets (i.e. {a,b},{b,c}, {b,d}) are calculated and values are shown in Table.8, Table.9 and Table.10. This is worthy to mention here that just for the representation purpose for instance {a_1, b_1} written as $\{a,b\}_1$, and similarly for others.

Table 8. Indiscernibility for the combination of feature {a,b}

RD	$\{a, b\}_1$	$\{a, b\}_2$	$\{a, b\}_3$
{a,b}	{1,2}	{3,4}	{5,6,7}

Table.9. Indiscernibility for the combination of feature {b,c}

RD	$\{b, c\}_1$	$\{b, c\}_2$	$\{b, c\}_3$	$\{b, c\}_4$	$\{b, c\}_5$	$\{b, c\}_6$
{b, c}	{1,2}	{5}	{6}	{7}	{3}	{4}

Table.10. Indiscernibility for the combination of feature {b,d}

RD	$\{b, d\}_1$	$\{b, d\}_2$	$\{b, d\}_3$	$\{b, d\}_4$	$\{b, d\}_5$	$\{b, d\}_6$
{b, d}	{1}	{2}	{5,6}	{7}	{3}	{4}

Similarly, the average conditional entropy values for the subsets {b,c} and {b,d} get calculated. The results are shown in Table 11.

Table.11. Conditional entropy values for the subsets {a, b}, {b, c} and {b, d}

Feature subset	{a, b}	{b, c}	{b, d}
Average Entropy	0.41105	0.07142	0

From the Table.11, it can be observed that feature subset {b,d} has attained the entropy value as 0, thus, the algorithm will stop and terminate by returning the reduct set RD as {b,d}.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

After going through, the detailed descriptions of the proposed methodology, this segment of the article elaborates the experimental results and analysis of the quantitative results which are obtained in this investigation.

4.1 EXPERIMENTAL SETUP

All the functions and programs are implemented in JAVA and Python 3.7 with Jupyter notebook environment. For simulation, the configuration of the computer system used as processor: Intel® Corei-5, 2.5 GHz clock speed, primary memory of 8GB and Operating system with Windows 10 environment. Details of the datasets used in this research are highlighted in the next Subsection. Individual medical/healthcare and life science related datasets are formatted in two ways; in the first format the class labels are considered and those are passed to the supervised SQR and EBR. In the other format it does not have the class labels and those are passed to URR, USQR, LSFS and RE-UFS methods to select different subsets of features. Because the SQR and EBR methods are fall under the supervised FS model, whereas URR, USQR, and LSFS falls under the unsupervised FS model.

For replacing the missing values by applying averaged k-NN method in several datasets namely Dermatology, Lung Cancer, Hepatitis, and Arrhythmia the range of k is considered as odd numbers [23]. The machine learning tool WEKA [32] is used for the purpose of classification. The WEKA tool is a popular open-source java-based machine-learning workbench. Very well-known two classifiers viz., SVM and RF are used in this investigation to predict the absence or presence of diseases**.** The classification accuracy with other quantitative measures of individual classifiers is obtained by computing the average of 15 times 10-fold cross validation (FCV) [1] technique by selecting different seed points in the WEKA platform.

Table 12: Details of Datasets

Dataset	(|U|)	(|C|)	(|D|)	Missing value
Diagnostic Wisconsin Breast Cancer Database (WDBC)	699	30	N/A	N/A
Dermatology	366	33	Yes	8
Lung Cancer	73	325	Yes	5
Arrhythmia	452	278	Yes	1180
Hepatitis	155	20	Yes	168
Cardiotocography	2126	22	N/A	N/A
Musk	476	167	N/A	N/A

Table.13: Number of features selected by various methods.

Dataset	SQR	EBR	URR	USQR	LSFS	RE-UFS
Diagnostic Wisconsin Breast Cancer Database (WDBC)	12	11	9	10	14	11
Dermatology	10	13	12	11	13	9
Lung Cancer	23	20	24	25	26	24
Arrhythmia	28	31	29	28	23	27
Hepatitis	8	9	8	8	9	8
Cardiotocography	13	11	14	13	14	12
Musk	19	22	20	21	24	23

4.2 DETAILS OF THE DATASETS

Different publicly available medical/healthcare domain benchmark datasets for the diseases related heart, cancer, skin and liver are collected from the UCI Machine Learning repositories [33] and ASU feature selection repository. But it is interesting to note that many of the datasets contains missing values in it and those are represented by '?' mark in the respective dataset. In Table 12, details of the seven (7) datasets are reported. Column 1, column 2, column 3, column 4, and column 5 represents the serial numbers, dataset name, number of objects (|U|) it contains, number of features (|C|), number of classes (|D|) and the number of missing values respectively.

Table.14. Experimental results with respect to accuracy, precision, recall, F-measure, MCC

Dataset	Method	SVM					RF				
		Acc (%)	Pre	Rec	F-M	MCC	Acc (%)	Pre	Rec	F-M	MCC
WDBC	SQR	94.79	0.917	0.919	0.880	0.866	93.85	0.950	0.940	0.910	0.911
	URR	92.13	0.903	0.921	0.845	0.782	90.27	0.944	0.933	0.915	0.915
	EBR	93.70	0.888	0.892	0.833	0.852	93.85	0.948	0.921	0.913	0.882
	USQR	94.79	0.912	0.910	0.66	0.845	93.35	0.910	0.922	0.920	0.890
	LSFS	91.22	0.822	0.820	0.711	0.776	90.11	0.829	0.865	0.855	0.872
	RE-UFS	**96.88**	**0.962**	**0.960**	**0.904**	**0.916**	**95.42**	**0.966**	**0.969**	**0.935**	**0.920**
Dermatology	SQR	93.60	0.822	0.825	0.775	0.766	91.24	0.837	0.833	0.902	0.776
	URR	91.69	0.824	0.817	0.708	0.782	89.50	0.894	0.890	0.887	0.725
	EBR	93.15	0.836	0.833	0.805	0.810	90.33	0.843	0.877	0.905	0.810
	USQR	89.23	0.766	0.772	0.778	0.566	90.70	0.834	0.859	0.844	0.822
	LSFS	90.66	0.778	0.750	0.764	0.773	91.55	0.823	0.833	0.778	0.803
	RE-UFS	**95.04**	**0.946**	**0.949**	**0.886**	**0.911**	**94.23**	**0.892**	**0.891**	**0.915**	**0.878**
Lung Cancer	SQR	61.87	0.646	0.619	0.664	0.770	61.25	0.647	0.613	0.602	0.503
	URR	65.01	0.631	0.651	0.589	0.633	62.50	0.636	0.520	0.588	0.422
	EBR	61.25	0.670	0.668	0.605	0.704	60.62	0.677	0.506	0.574	0.466
	USQR	65.87	0.615	0.614	0.623	0.865	67.25	0.703	0.663	0.610	0.592
	LSFS	66.55	0.744	0.656	0.522	0.776	62.77	0.722	0.688	0.604	0.584
	RE-UFS	**69.87**	**0.711**	**0.751**	**0.606**	**0.865**	**72.50**	**0.740**	**0.715**	**0.655**	**0.610**
Arrhythmia	SQR	69.88	0.682	0.614	0.662	0.688	68.05	0.650	0.602	0.683	0.533
	URR	65.18	0.642	0.519	0.610	0.588	65.20	0.629	0.629	0.634	0.597
	EBR	68.87	0.686	0.640	0.538	0.633	67.88	0.595	0.593	0.602	0.422
	USQR	68.07	0.697	0.549	0.620	0.610	66.07	0.592	0.602	0.622	0.488
	LSFS	64.66	0.744	0.566	0.595	0.655	69.03	0.633	0.632	0.586	0.407
	RE-UFS	**71.23**	**0.769**	**0.716**	**0.670**	**0.703**	**69.81**	**0.673**	**0.693**	**0.690**	**0.588**
Hepatitis	SQR	93.16	0.905	0.910	0.877	0.895	94.12	0.892	0.912	0.904	0.880
	URR	85.11	0.810	0.815	0.855	0.820	87.12	0.722	0.721	0.766	0.791
	EBR	88.40	0.777	0.780	0.766	0.840	93.12	0.880	0.892	0.866	0.855
	USQR	91.26	0.888	0.894	0.805	0.910	95.12	0.901	0.901	0.887	0.910
	LSFS	**95.33**	**0.944**	**0.930**	**0.905**	**0.955**	**97.19**	**0.959**	**0.952**	**0.922**	**0.930**
	RE-UFS	**95.33**	**0.944**	**0.930**	**0.905**	**0.955**	**97.19**	**0.959**	**0.952**	**0.922**	**0.930**
Cardiotocography	SQR	91.19	0.792	0.813	0.907	0.773	92.88	0.805	0.819	0.804	0.788
	URR	87.22	0.821	0.780	0.823	0.655	86.44	0.801	0.716	0.766	0.767
	EBR	90.57	0.881	0.866	0.792	0.689	88.70	0.780	0.786	0.755	0.739
	USQR	88.19	0.887	0.810	0.797	0.744	89.30	0.886	0.810	0.788	0.824
	LSFS	90.55	0.866	0.902	0.855	0.758	91.33	0.908	0.907	0.866	0.822
	RE-UFS	**93.20**	**0.887**	**0.810**	**0.915**	**0.744**	**89.30**	**0.886**	**0.810**	**0.877**	**0.844**
Musk	SQR	**90.28**	**0.890**	**0.888**	**0.855**	**0.695**	**92.17**	**0.805**	**0.819**	**0.804**	**0.709**
	URR	86.23	0.833	0.830	0.803	0.602	83.27	0.711	0.706	0.655	0.655
	EBR	79.50	0.788	0.765	0.752	0.456	82.06	0.733	0.711	0.588	0.612
	USQR	81.15	0.707	0.689	0.79	0.554	83.50	0.765	0.766	0.535	0.588
	LSFS	82.95	0.755	0.677	0.855	0.528	81.33	0.718	0.720	0.703	0.633
	RE-UFS	88.20	0.845	0.844	0.915	0.658	90.30	0.775	0.791	0.722	0.682

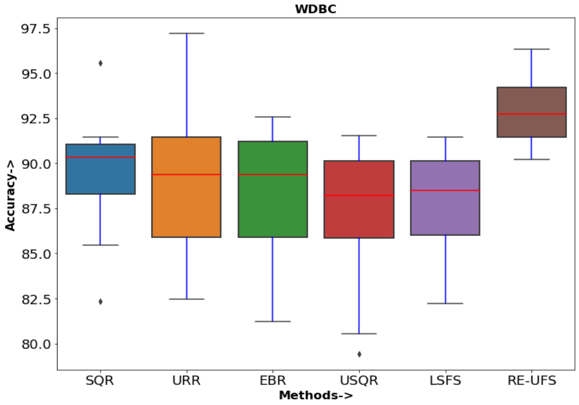
4.3 CLASSIFIER PERFORMANCE EVALUATION MEASURES

Five different classification validity measures [1] [22] such as (i) average accuracy, (ii) average precision (Pre), (iii) average recall (Rec) (iv) F-score measure (F-M) and (v) Matthew's correlation coefficient (MCC) is used to check the effectiveness of the proposed RE-UFS and compared methods with respect to two classifiers namely SVM and RF. Accuracy is the total number of correct predictions divided by the total number of predictions made for a dataset. Precision is the ratio between the True Positives and all the Positives. Recall is the measure of a model that correctly identify the True Positives. F-M provides a way to combine both precision and recall into a single measure.

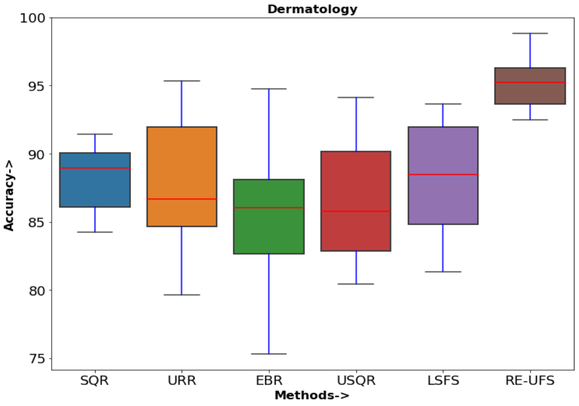
4.4 RESULT DISCUSSIONS

In this section the detailed descriptions of the quantitative analyses of the experimental results are given. The performances of the proposed method and other compared methods are also demonstrated. Based on the results obtained by the FS methods, Table.13 summarized the number of features selected by the various methods. Then each reduct sets generated by individual methods for different datasets are passed to the classifiers to evaluate the performance of the FS algorithms. The experimental results achieved by applying the average of the 15 times 10-FCV in WEKA environment, are reported in Table.14.

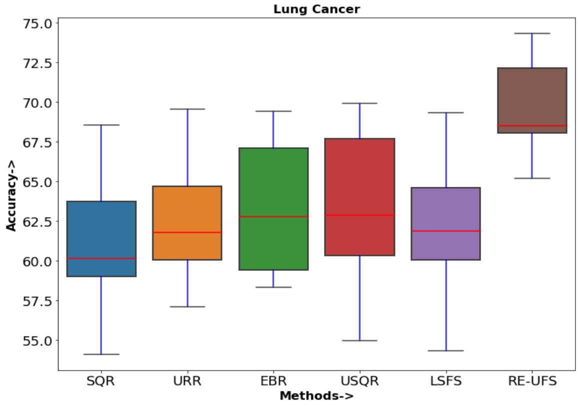
From the tabulated results of Table.14 for the selected features and experimental quantitative results of 10-FCV it can be observed that for the WDBC dataset, proposed RE-UFS method performed better than all the other compared methods. The classifier validity measures such as accuracy, precision, recall, F-measure, and MCC for the SVM and RF classifiers are better compared to other methods. The improvements of classification accuracy for the proposed method obtained by the two classifiers namely SVM and RF are [{2.09%, 4.75%, 3.18%, 2.09%, and 0.3.57%}, {1.57%, 5.15%, 1.57%, 2.07% and 5.31%}] with respect to SQR, URR, EBR, USQR and LSFS methods. For the Dermatology dataset, the better classification accuracy obtained by the RE-UFS method by the two classifiers viz., SVM and RF are [{1.44%, 3.35%, 1.89%, 5.81% and 4.38%}, {2.99%, 4.73%, 3.90%, 3.53% and 2.68%}] with respect to the SQR, URR, EBR, USQR and LSFS methods.



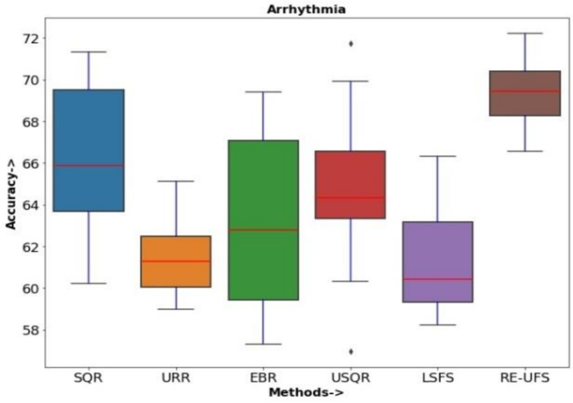
(a)



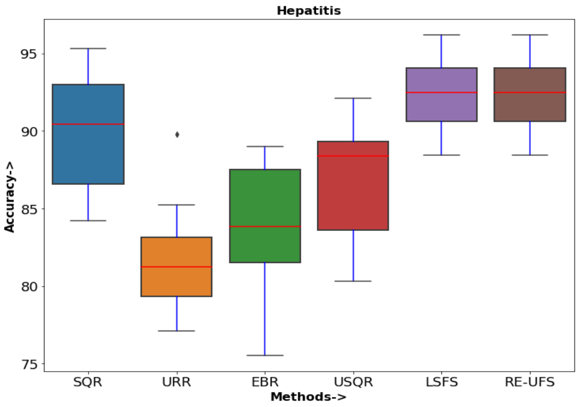
(b)



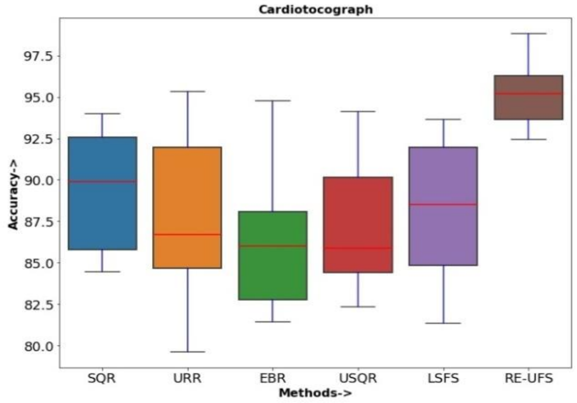
(c)



(d)



(e)



(f)

Fig.3. Summary of Boxplots from (a-f) describing the average of classification accuracy achieved by the RF and SVM classifiers with respect to the six different FS methods on six datasets

Similarly, for the lung cancer (discrete) dataset, it can be observed from the Table.14 that RE-UFS method performed better than all the other compared methods in terms of all the validity measures for the SVM and RF classifiers results. Moving onto the Arrhythmia dataset, which is one of the complex datasets among the other considered datasets in this research. The number of missing values in this dataset is very dense. By considering the value of (k=9) data k-NN imputation method is applied in this dataset. Respective classification accuracy improvements produced by SVM and RF with respect to SQR, URR, EBR, USQR and LSFS are [{1.35%, 6.05%, 2.36%, 3.16% and 6.57%}, {1.76%, 4.61%, 1.93%, 3.74% and 0.78%}. It is interesting to note that there is no such difference in the results of Hepatitis data for LSFS and RE-UFS methods. But compared to other four RST based models, RE-UFS method has dominates the results. Similarly, for the Cardiotocography dataset, for the selected features of RE-UFS method, SVM and RF classifiers gives better results compared to other counterpart methods. However, for the selected features of SQR method, both the classifiers have achieved better results compared to RE-UFS method, for the Musk dataset. To predict the new molecules as musks or non-musks, SQR method has select better features. Moreover, Boxplots [13] of average classification accuracies of FCV simulations obtained by the two different classifiers (RF and SVM) on the six datasets namely WDBC, dermatology, lung cancer, arrhythmia, hepatitis and cardiotocography are shown in Fig.3. The figures justify the higher median values of the average accuracies produced by the RF and SVM for the proposed RE-UFS method in comparison to the other FS methods. Thus, these Boxplots indicates the fact that the classification accuracies achieved by the classifiers for the proposed RE-UFS method are better than the other compared methods. Thus, after comparing the experimental results, the effectiveness of the proposed RE-UFS method with respect to the five compared method results, a conclusion can be drawn that the performance of the RE-UFS algorithm is better than the performance of the SQR, EBR, URR, USQR, and LSFS algorithms. By employing the RST with information theory framework, we have successfully reduced the number features without compromising on the accuracy of the classifier. RF and SVM both the classifiers turned out to be the best machine learning model to classify the medical/healthcare datasets.

5. CONCLUSIONS AND FUTURE WORK

FS is one of the essential steps in data mining technique. Moreover, it is an important step in many of the latest application domains of machine learning. The continuous increase of the average dataset sizes in different domains, such as engineering and medical sciences, demands this FS process. FS methods can boost the performance of machine learning algorithms by reducing the complexity and time of execution. In this article, an unsupervised FS method underlying the theory of rough set theory with an information theory model for the task of classification of different medical/healthcare and life science datasets is proposed. The significance of the proposed unsupervised RE-UFS method is verified on seven real-life medical and life science-related datasets. Then, to validate, the proposed method's subset selected features are passed to two well-known classifiers, SVM and RF. Five different validity measures viz., accuracy, precision, recall, F-measure and MCC are calculated. Additionally, Boxplot representation concerning the six datasets average classification is also evaluated.

The proposed method is compared with four well known existing rough set-based FS methods and with one unsupervised algorithm for the classification task. It is verified from the experimental section results that the proposed RE-UFS method has effectively performed superior in terms of all the validity measures for the six (6) datasets in comparison to other five state-of-the-art methods. Hence, the proposed RST-based unsupervised FS approach can be helpful in the diagnosis of the disease over the different existing approaches.

In the future, the proposed unsupervised FS method can also be effectively tested on other real-time high-dimensional datasets (text mining, genetics, or bioinformatics) with some enhancement or considerable modifications. Moreover, by choosing different machine learning algorithms (classifiers) alongside the FS techniques may alter the outcome which will be investigated in our future work.

REFERENCES

[1] J. Han and M. Kamber, "*Data Mining Concepts and Techniques*", Morgan Kaufmann Publishers, 2012.

[2] J.P. Cunningham and Z. Ghahramani, "Linear Dimensionality Reduction: Survey, Insights, and Generalizations", *Journal of Machine Learning Research*, Vol. 16, No. 1, pp. 2859-2900, 2018.

[3] R.K. Bania, "Survey on Feature Selection for Data Reduction", *International Journal of Computer Applications*, Vol. 94, No. 18, pp. 1-7, 2014.

[4] D. Jain and V. Singh, "Feature Selection and Classification Systems for Chronic Disease Prediction: A Review", *Egyptian Informatics Journal*, Vol. 19, No. 3, pp.179-189, 2018.

[5] L. Wolf and A. Shashua, "Feature Selection for Unsupervised and Supervised Inference: The Emergence of Sparsity in a Weight-Based Approach", *Journal of Machine Learning Research*, Vol. 6, pp.1855-1887, 2005.

[6] X. Yan, B. Gebru and E. Tunstel, "An Efficient Unsupervised Feature Selection Procedure through Feature Clustering", *Pattern Recognition Letters*, Vol. 131, pp. 277-284, 2020.

[7] X. He and P. Niyogi, "Laplacian Score for Feature Selection", *Proceedings of International Conference on Advances Neural Information Processing*, pp. 507-514, 2005.

[8] C. Bancioiu and L. Vintan, "Efficiency Optimizations for Koller and Sahami's Feature Selection Algorithm", *Romanian Journal of Information Science and Technology*, Vol. 22, No. 1, pp. 85-99, 2019.

[9] L. Sun and X. Cao, "Decision Table Reduction Method Based on New Conditional Entropy for Rough Set Theory", *International Workshop on Intelligent Systems and Applications*, Vol. 25, pp.759-768, 2009.

[10] J. Liang, C. Dang and Y. Qian, "An Efficient Rough Feature Selection Algorithm with a Multi-Granulation View", *International Journal of Approximate Reasoning*, Vol. 53, pp.912-926, 2012.

[11] A. Chouchoulas and Q. Shen, "Rough Set-Aided Keyword Reduction for Text Categorization", *Applied Artificial Intelligence: An International Journal*, Vol. 15, No. 9, pp. 843-873, 2003.

[12] H.H. Inbarani, A.T. Azar and G. Jothi, "Supervised Hybrid Feature Selection based on PSO and Rough Sets for Medical Diagnosis", *Computer Methods and Programs in Biomedicine*, Vol. 113, pp. 175-185, 2014.

[13] A. Arshaghi, M. Ashourian and L. Ghabeli, "Detection of Skin Cancer Image by Feature Selection Methods Using New Buzzard Optimization (BUZO) Algorithm", *Traitement Du Signal*, Vol. 37, No. 2, pp. 181-194, 2020.

[14] C. Velayutham and K. Thangavel, "Rough Set based Unsupervised Feature Selection using Relative Dependency Measures", *International Journal of Computational Intelligence and Informatics*, Vol. 1, No. 1, pp. 64-69, 2011.

[15] K. Thangavel, "Unsupervised Quick Reduct Algorithm using Rough Set Theory", *Journal of Electronic Science and Technology*, Vol. 9, No. 3, pp.193-201, 2011.

[16] V.B. Canedo and A. Betanzos, "A Review of Feature Selection Methods on Synthetic Data", *Knowledge Information System*, Vol. 34, pp. 483-519, 2013.

[17] S. Shilaskar and A. Ghatol, "Feature Selection for Medical Diagnosis: Evaluation for Cardiovascular Diseases", *Expert Systems with Applications*, Vol. 40, pp. 4146-4153, 2013.

[18] P.K.N. Banu and H.H. Inbarani, "Rough Set Based Feature Selection for Egyptian Neonatal Jaundice", *Proceedings of International Conference on Advanced Machine Learning Technologies and Applications*, pp. 367-378, 2014.

[19] G. Jothi and H. Inbarani, "Soft Set Based Quick Reduct Approach for Unsupervised Feature Selection", *Proceedings of IEEE International Conference on Advanced Communication Control and Computing Technologies*, pp. 277-281, 2012.

[20] E.S. Shamery and A.R. Al-Obaidi, "A New Approach of Rough Set Theory for Feature Selection and Bayes Net Classifier Applied on Heart Disease Dataset", *Journal of Babylon University Pure and Applied Sciences*, Vol. 26, No. 2, pp. 15-26, 2018.

[21] Y. Wang and L. Ma, "Feature Selection for Medical Dataset using Rough Set Theory", *Proceedings of IEEE International Conference on Computer Engineering and Applications*, pp. 68-72, 2009.

[22] R.K. Bania and R. Halder, "R-Ensembler: A Greedy Rough Set based Ensemble Attribute Selection, Algorithm with K-NN Imputation for Classification of Medical Data", *Computer Methods and Programs, in Biomedicine*, Vol. 184, pp. 105122-105133, 2020.

[23] J. Chen and J. Shao, "Nearest Neighbor Imputation for Survey Data", *Journal of Official Statistics*, Vol. 16, No. 2, pp. 113-131, 2000.

[24] A. Farhangfar and J. Dy. "Impact of Imputation of Missing Values on Classification Error for Discrete Data", *Pattern Recognition*, Vol. 41, pp. 3692-3705, 2008.

[25] P. Schmitt, J. Mandel and M. Guedj, "A Comparison of Six Methods for Missing Data Imputation", *Journal of Biometrics and Biostatistics*, Vol. 6, No. 1, pp. 1-6, 2015.

[26] K.B. Nahato, K.N. Harichandran and K. Arputhara, "Knowledge Mining from Clinical Datasets using Rough Sets and Backpropagation Neural Network", *Proceedings of IEEE International Conference on Computational and Mathematical Methods in Medicine*, pp. 1-3, 2015.

[27] P. Yildirim, "Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease", *International Journal of Machine Learning and Computing*, Vol. 5, No. 4, pp. 258-263, 2015.

[28] H. Liu and M. Dash, "Discretization: An Enabling Technique", *Data Mining and Knowledge Discovery*, Vol. 6, pp. 393-423, 2002.

[29] C.J. Tsai and W.P. Yang, "A Discretization Algorithm based on Class-Attribute Contingency Coefficient", *Information Sciences*, Vol. 178, pp.714-731, 2008.

[30] Y. Wei, T. Liu and R. Valdez, "Application of Support Vector Machine Modeling for Prediction of Common Diseases: The Case of Diabetes and Pre-Diabetes", *BMC Medical Informatics and Decision Making*, Vol. 10, No. 16, 2020.

[31] Y. Yang and W. Cai. "Using Random Forest for Reliable Classification and Cost-Sensitive Learning for Medical Diagnosis", *BMC Bioinformatics*, Vol. 10, No. 1, pp. 1-14, 2009.

[32] Weka Machine Learning Tool, Available at https://www.cs.waikato.ac.nz/ml/ weka.html, Accessed at 2021.

[33] C.L. Blake, "UCI Repository of Machine Learning Databases", Available at https://www.ics.uci.edu/~mlearn, Accessed at 2022.

.