Volume 4, Issue 5, May 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Application of Robust Nearest Neighbour Fuzzy Rough Set (RNN-FRS) to Feature Selection

Bichitrananda Behera
Department of Computer Science and Engineering,
Biju Patnaik University of Technology, India

Abstract— Goal of feature selection is to omit features (attributes) from decision systems such that objects in different decision classes can still be discerned. The fuzzy dependency function proposed in fuzzy rough set is widely employed in feature evaluation and attribute reduction. It was shown earlier that this function is not robust to noise information. As datasets in real-world applications are usually contaminated by noise, robustness of data analysis models is very important in practice. Therefore, in this paper, it is considered a more flexible methodology based on recently introduced Robust Nearest Neighbour Fuzzy Rough Set (RNN-FRS) model. Then the fuzzy dependency function of RNN-FRS is used to evaluate and select features. The presented experimental results show the effectiveness of the RNN-FRS based feature selection method (FRS-FS).

Keywords— Feature selection, fuzzy rough set, robustness, rough set, robust nearest neighbour

I. INTRODUCTION

In classification learning, data are usually described with a great number of features. Typically, some parts are irrelevant or redundant with the classification task. These irrelevant features might confuse learning algorithms and deteriorate learning performance. Hence, it is useful to select relevant and indispensable features for designing classification systems. So far, a number of algorithms have been developed for feature reduction [6,7,8,9,10,11] . Generally speaking, there are two key issues in constructing a feature selection algorithm: feature evaluation and search strategies. Feature evaluation is used to measure the quality of the candidate features. Obviously, evaluation functions have great influence on outputs of algorithms. A great number of functions were designed, such as dependency [12], neighbourhood dependency [13] and fuzzy dependency in the rough set theory[14,15]; mutual information and symmetric uncertainty in information theory[11];sample margin [16] and hypothesis margin [17,18] in statistical learning theory, and so on. As to the search strategy, it can be roughly divided into two categories. One guarantees to find the optimal subset of features in terms of the used evaluation function, such as the exhaustive search [19] and the branch-and-bound algorithm [22]. And the other is to find a suboptimal solution for efficiency, including sequential forward selection [20], sequential backward elimination [19], floating search [21,22], mRMR [23], etc.

The rough set theory provides a mathematical tool to handle uncertainty in data analysis [1]. It has been successfully used in attribute reduction and rule learning [2,3]. Moreover, this theory also provides practical solutions to many data analysis tasks, such as data mining and rule discovery. The classic rough set model is defined with equivalence relations, which leads to the limitation in handling data with numerical or fuzzy attributes, some generalized models were proposed, such as fuzzy rough sets [4,5].

It is well known that datasets in real-world application are usually corrupted by noise. The noisy samples may have great influence on outputs of the models. Accordingly, the performance of classification systems would be reduced. So, robust models and algorithms are highly desirable in practice.

In the framework of rough sets, dependency functions, defined as the ratio of the consistent samples over the universe, are used to compute the quality of features. This function plays the central role in rough set based learning algorithms. However, it is observed that the dependency function defined in Pawlak rough set model is not robust. This property is passed down to neighbourhood rough sets and fuzzy rough sets, which limits the applications of these models.

In order to deal with this problem, some extended models were developed. First, Yao, Wong et al. proposed the decision theoretic rough set model (DTRS) in 1990 [24] and applied this model to attribute reduction in 2008 [25]. This model considers the statistic information in data. In 1993,Ziarko developed the variable precision rough set model (VPRS) to tolerate noisy samples [26], where several mislabel samples in an equivalence class are overlooked in computing lower and upper approximations. However, given a learning task, it is a big problem to set how many samples should be overlooked. In addition, information theory was also introduced to compute the significance of features [27]. These models are indeed more robust than rough sets, however, the granular structures are lost in these models. In [28], a comparative study between Pawlak’s rough sets based reduction and the information-theoretic based reduction was conducted.

In addition, in 2007,Cornelis *et al.* presented a model called vaguely quantified rough sets (VQRS) [29], which was used in constructing a robust feature selection algorithm in 2008 [30]. In 2009,Rolka et al. and Zhao, Tsang et al. showed the definitions of variable precision fuzzy rough sets and fuzzy variable precision rough sets[31] to enhance robustness of fuzzy rough sets, respectively and apply it to feature selection. Unfortunately, we find that the model in [31] is still sensitive to mislabel samples. Although there are some models to deal with noise in datasets, it seems that handling noise is still an open problem in the rough set theory.

In 2010,Hu *et al.* introduced a new robust model of fuzzy rough sets, which are called soft fuzzy rough sets, where soft threshold was used to compute fuzzy lower and upper approximations and feature selection was used based on that soft fuzzy rough set [32].In 2012,Hu et al. introduced robust nearest neighbour fuzzy rough set (RNN-FRS) model and compared RNN-FRS model with other robust models like β -PFRS, VQRS, VPFRS, FVPRS and SFRS and it performed best among them[33].

According to my knowledge there was no extensive work have been done to apply robust nearest neighbour fuzzy rough set (RNN-FRS) to feature selection. In this paper RNN-FRS is used to feature selection and it is compared with fuzzy rough set (FRS) based feature selection. As we know Robust Nearest Neighbour Fuzzy Rough Classifier (RNN-FRC) is a best classifier [35]. So, here RNN-FRC is used as classifier. Some numerical experiments have been done to show RNN-FRS based feature selection performs better than fuzzy rough set based feature selection.

The remainder of this paper is organized as follows. First, preliminary knowledge of feature selection, rough sets and fuzzy rough sets is given in Section II; then, the existing models of robust nearest neighbour fuzzy rough set models (RNN-FRS) is discussed in Section III. Then, robust nearest neighbour fuzzy rough set based feature selection (RNN-FS) algorithm is explained in section IV. Experimental analysis is given in Section V. Finally, conclusions and future scope are drawn in Section VI.

II. Preliminaries

In this section basic concept of feature selection, rough set and fuzzy rough set are discussed.

A. Feature Selection

Feature selection (FS) or Attribute Selection (AS) refers to the problem of selecting those input attributes or features that are most predictive of a given outcome; a problem encountered in many areas such as machine learning, pattern recognition and signal processing. Unlike other dimensionality reduction methods, feature selectors preserve the original meaning of the attributes after reduction. This has found application in tasks that involve datasets containing huge numbers of attributes (in the order of tens of thousands) which, for some learning algorithms, might be impossible to process further. Recent examples include text processing and web content classification. FS techniques have also been applied to small and medium sized datasets in order to locate the most informative attributes for later use. In this paper, the dependency function is RNN-FD.

FS techniques attempt to retain the meaning of the original attribute set. The main aim of attribute selection is to determine a minimal attribute subset from a problem domain while retaining a suitably high accuracy in representing the original attributes. In many real world problems, as is a must due to the abundance of noisy, irrelevant or misleading attributes.

The usefulness of an attribute or attribute subset is determined by both its *relevancy* and *redundancy*. An attribute is said to be relevant if it is predictive of the decision attribute(s), otherwise it is irrelevant. An attribute is considered to be redundant if it is highly correlated with other attributes. Hence, the search for a good attribute subset involves finding those attributes that are highly correlated with the decision attribute(s), but are uncorrelated with each other

Given an attribute set size, the task of FS can be seen as a search for an “optimal” attribute subset through the competing candidate subsets. The definition of what an optimal subset is may vary depending on the problem to be solved. Although an exhaustive method may be used for this purpose in theory, this is quite impractical for most datasets. Usually AS algorithms involve heuristic or random search strategies in an attempt to avoid this prohibitive complexity. However, the degree of optimality of the final attribute subset is often reduced. The overall procedure for any attribute selection method is given in Fig. 1[14].

The generation procedure implements a search method that generates subsets of attributes for evaluation. It may start with no attributes, all attributes, a selected attribute set or some random attribute subset. Those methods that start with an initial subset usually select these attributes heuristically beforehand. Attributes are added (*forward selection*) or removed (*backward elimination*) iteratively in the first two cases [14]. In the last case, attributes are either iteratively added or removed or produced randomly thereafter. An alternative selection strategy is to select instances and examine differences in their attributes. The evaluation function calculates the suitability of an attribute subset produced by the generation procedure and compares this with the previous best candidate, replacing it if found to be better.

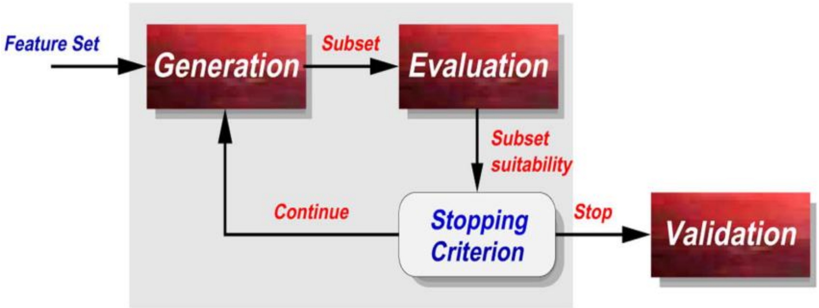


Figure1. Feature selection process [14].

A stopping criterion is tested every iteration to determine whether the FS process should continue or not. For example, such a criterion may be to halt the FS process when a certain number of attributes have been selected if based on the generation process. A typical stopping criterion centred on the evaluation procedure is to halt the process when an optimal subset is reached. Once the stopping criterion has been satisfied, the loop terminates. For use, the resulting subset of attributes may be validated. Determining subset optimality is a challenging problem. There is always a trade-off in non-exhaustive techniques between subset minimal and subset suitability—the task is to decide which of these must suffer in order to benefit the other. For some domains (particularly where it is costly or impractical to monitor many attributes), it is much more desirable to have a smaller, less accurate attribute subset. In other areas it may be the case that the modelling accuracy (e.g., the classification rate) using the selected attributes must be extremely high, at the expense of a non-minimal set of attributes.

Feature selection algorithms may be classified into two categories based on their evaluation procedure. If an algorithm performs FS independently of any learning algorithm (i.e., it is a completely separate pre-processor), then it is a *filter* approach. In effect, irrelevant attributes are filtered out before induction. Filters tend to be applicable to most domains as they are not tied to any particular induction algorithm. If the evaluation procedure is tied to the task (e.g., classification) of the learning algorithm, the FS algorithm employs the *wrapper* approach. This method searches through the attribute subset space using the estimated accuracy from an induction algorithm. As a measure of subset suitability, although wrappers may produce better results, they are expensive to run and can break down with very large numbers of attributes. This is due to the use of learning algorithms in the evaluation of subsets, some of which can encounter problems when dealing with large datasets.

B. Rough Set

The Rough set concept can be defined quite generally by means of topological operations, interior and closure, called approximations. $IS = \langle \mathrm{U}, \mathrm{A} \rangle$ is called an information table, where U is a finite and nonempty set of objects and A is a set of features used to characterize the objects. $\forall \mathrm{B} \subseteq \mathrm{A}$, a B-indiscernibility relation is defined as

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

Then the partition of U generated by IND (B) is denoted by U/IND(B) (or U/B). The equivalence class of x induced by B-indiscernible relation is denoted by $[x]_B$.

Given an arbitrary $\mathrm{X} \subseteq \mathrm{U}$, R is an equivalence relation on U induced by a set of attributes.

- The lower approximations of X with respect to R are defined as

$$\underline{R}X = \{\mathrm{x} \in \mathrm{U} \mid [\mathrm{x}]_R \subseteq \mathrm{X}\}$$

The lower approximation of a set X with respect to R is the set of all objects, which can be for certain classified as X with respect to R (are certainly X with respect to R)

- The upper approximations of X with respect to R are defined as

$$\overline{R}X = \{\mathrm{x} \in \mathrm{U} \mid [\mathrm{x}]_R \cap \mathrm{X} \neq \phi\}$$

The upper approximation of a set X with respect to R is the set of all objects which can be possibly classified as X with respect to R (are possibly X in view of R).

- R-boundary region of X is defined as

$$\mathrm{BN}_R(\mathrm{X}) = \overline{R}X - \underline{R}X$$

The boundary region of a set X with respect to R is the set of all objects, which can be classified neither as X nor as not-X with respect to R.

- R-negative region of X is defined as

$$NEG_R(\mathrm{X}) = \mathrm{U} - \overline{R}X$$

It contains those elements which completely do not belong to X.

The definition of rough sets is

- Set X is crisp (exact with respect to R), if the boundary region of X is empty.
- Set X is rough (inexact with respect to R), if the boundary region of X is nonempty

The lower approximation is also called R-positive region of X, denoted by $POS_R(\mathrm{X})$ .Given a decision table $\mathrm{DS} = \langle U, A \cup D \rangle$ D is the decision attribute. For $\forall \mathrm{B} \subseteq \mathrm{A}$, the positive region of decision D on B, denoted by $POS_B(\mathrm{D})$, is defined as

$$POS_B(\mathrm{D}) = \bigcup_{X \in U/D} \underline{B}X \qquad (1)$$

Where, U/D is the set of the equivalence classes generated by D. The dependency of decision D on B is defined as

$$\gamma_B(\mathrm{D}) = \frac{|POS_B(\mathrm{D})|}{|U|} \qquad (2)$$

Dependency is the ratio of the samples in the lower approximation over the universe. As the lower approximation is the set of objects with consistent decisions, dependency is used to measure the classification performance of attributes. It is expected that all the decisions of objects are consistent with respect to the given attributes. In practice, inconsistency widely exists in data.

C. Fuzzy Rough Set

The Rough set model is constructed under the assumption that only discrete features exist in the information system. In practice, most of classification tasks are described with numerical features or fuzzy information. In this case, neighbourhood relations or fuzzy similarity relations are used and neighbourhood or fuzzy granules are generated. Then, we use these granules to approximate decision classes.

Given a nonempty universe *U*, *R* is a fuzzy binary relation on *U*. If *R* satisfies

(1) Reflexivity: $R(x, x) = 1$
(2) Symmetry: $R(x, y) = R(y, x)$
(3)Sup-min transitivity: $R(x, y)\ \sup\min_{z\in U}\{\mathrm{R(x,z),R(z,y)}\}$

Then *R* is a fuzzy equivalence relation. The fuzzy equivalence class $[x]_{\mathrm{R}}$ associated with *x* and *R* is a fuzzy set on *U*, where $[\mathrm{x}]_R(\mathrm{y}) = \mathrm{R(x,y)}$ for all *y ϵU.*

Let *U* be a nonempty universe, *R* be a fuzzy equivalence relation on *U* and *X* (*U*) be the fuzzy power set of *U*. Given a fuzzy set $X \in X(\mathrm{U})$, the lower and upper approximations are defined as

$$\begin{cases} \underline{R}X(\mathrm{x}) = \inf\limits_{y\in U} \max(1-\mathrm{R(x,y)},\mathrm{X(y)}) \\ \overline{R}X(\mathrm{x}) = \sup\limits_{y\in U} \min(\mathrm{R(x,y)},\mathrm{X(y)}) \end{cases} \tag{3}$$

These approximation operators were discussed in the view point of the constructive and axiomatic approaches. In 1998, Morsi and Yakout replaced fuzzy equivalence relation with a *T*-equivalence relation and built an axiom system of the model, where the lower and upper approximations of *X* ∈ *X* (*U*) are

$$\begin{cases} \underline{Rs}X(\mathrm{x}) = \inf\limits_{y\in U}\mathrm{S(N(R(x,y)),X(y))} \\ \overline{R_T}X(\mathrm{x}) = \sup\limits_{y\in U}\mathrm{T(R(x,y),X(y))} \end{cases} \tag{4}$$

Where, *T* is a triangular norm.

In 2002,Radzikowska and Kerre introduced a model :

$$\begin{cases} \underline{R_\vartheta}X(\mathrm{x}) = \inf\limits_{y\in U}\vartheta(\mathrm{R(x,y),X(y)}) \\ \overline{R_\sigma}X(\mathrm{x}) = \sup\limits_{y\in U}\sigma(\mathrm{N(R(x,y)),X(y)}) \end{cases} \tag{5}$$

In classification learning, samples are assigned with a class label and described with a group of features. Fuzzy equivalence relations can be generated with numerical or fuzzy features, while the decision variable divides the samples into some subsets. In this case, the task is to approximate these decision classes with the fuzzy equivalence classes induced with the features. Given a decision system <*U*, *R*, *D*>, for a decision class $D_i \in U/D$, the membership of a sample *x* to D_i is

$$D_i(\mathrm{x}) = \begin{cases} 1x \in D_i \\ 0x \notin Di \end{cases} \tag{6}$$

Then the membership of sample *x* to the fuzzy lower approximation of D_i is

$$\begin{aligned} \underline{R}D_i(\mathrm{x}) &= \inf_{y\in U} \max\ \{1-\mathrm{R(x,y)},\mathrm{D}_i(\mathrm{y})\} \\ &= \inf_{y\in D_i}\max\{1-\mathrm{R(x,y)},1\} \wedge \inf_{y\notin D_i}\max\{1-\mathrm{R(x,y)},0\} \\ &= 1\wedge \inf_{y\notin D_i}\{1-\mathrm{R(x,y)}\} = \inf_{y\notin D_i}\{1-\mathrm{R(x,y)}\} \end{aligned} \tag{7}$$

If the Gaussian kernel function is used to compute the similarity *R*,

$$G(\mathrm{x,y}) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{8}$$

1 − *G*(*x*, *y*) can be considered as a pseudo-distance function. Similarly, the membership of sample *x* to the fuzzy upper approximation of D_i is

$$\begin{aligned}\overline{R}D_i(\mathrm{x}) &= \sup_{y\in D_i}\min\{\mathrm{R(x,y)}, \mathrm{D}_i(\mathrm{y})\}\\ &= \sup_{y\in D_i}\min\{\mathrm{R(x,y)},1\}\vee\sup_{y\notin D_i}\min\{\mathrm{R(x,y)},0\}\\ &= \sup_{y\in D_i}\min\{\mathrm{R(x,y)}\}\vee 0 = \sup_{y\in D_i}\min\{\mathrm{R(x,y)}\}\end{aligned}\tag{9}$$

We can see that $\underline{R}D_i(\mathrm{x})$ is the distance from *x* to its nearest sample from different classes; while $\overline{R}D_i(\mathrm{x})$ is the similarity between *x* and the nearest sample in D_i .

III. Robust Nearest Neighbour Fuzzy Rough Set

It was shown that *both* $\underline{R_S}d_i(\mathrm{x})$ or $\overline{R_T}d_i(\mathrm{x})$ depends on the nearest miss of *x*, i.e., the nearest sample from different classes of *x*. As we know, the statistics of minimum and maximum are very sensitive to noisy samples. Just one noisy sample would change the minimum or maximum of a random variable. The sensitiveness of these statistics leads to the poor performance of fuzzy rough sets in dealing with noisy datasets. RNN-FRS introduced robust statistics to substitute the operators of minimum and maximum in the fuzzy rough set model. So that it can be performed better in noisy environment.

A. Basic Definitions:

Given a random variable *X* and its *n* samples $x_1, x_2, ..., x_n$ sorted in the ascending order, the *k*-trimmed minimum of *X* is x_{k+1}; the *k*-trimmed maximum of *X* is x_{n-k-1}; *k*-mean minimum of *X* is $\sum_{i=1}^{k} x_i / k$; *k*-mean maximum of *X* is $\sum_{i=n-k}^{n} x_i / k$, and *k*-median minimum of *X* is median($x_1, x_2, ..., x_k$); *k*-mean maximum of *X* is median($x_{n-k}, ..., x_n$), denoted by $\min_{k-trimmed}(\mathrm{X})$, $\max_{k-trimmed}(\mathrm{X})$, $\min_{k-mean}(\mathrm{X})$, $\max_{k-mean}(\mathrm{X})$, $\min_{k-median}(\mathrm{X})$, and $\max_{k-median}(\mathrm{X})$, respectively.

Given $DT = \langle U, A, D\rangle$, *R* is a fuzzy similarity relation induced by *B* is subset *C* and *R*(*x*, *y*) monotonously decreases with their distance $\|x - y\|$. If d_i is one class of samples labelled with *i* and $x \in d_i$, then the robust fuzzy rough operators are defined as

$$\begin{aligned}\underline{R_S}_{\,k-trimmed}d_i(\mathrm{x}) &= \min_{y\notin d_{k-trimmed}} 1 - R(\mathrm{x,y})\\ \overline{R_T}_{\,k-trimmed}d_i(\mathrm{x}) &= \max_{y\in d_{i_{k-trimmed}}} R(\mathrm{x,y})\\ \underline{R_\vartheta}_{\,k-trimmed}d_i(\mathrm{x}) &= \min_{y\notin d_{i_{k-trimmed}}} \sqrt{1-R^2(\mathrm{x,y})}\\ \overline{R_\sigma}_{\,k-trimmed}d_i(\mathrm{x}) &= \max_{y\in d_{i_{k-trimmed}}} 1-\sqrt{1-R^2(\mathrm{x,y})}\end{aligned}\tag{10}$$

$$\begin{aligned}\underline{R_S}_{\,k-mean}d_i(\mathrm{x}) &= \min_{y\notin d_{k-mean}} 1 - R(\mathrm{x,y})\\ \overline{R_T}_{\,k-mean}d_i(\mathrm{x}) &= \max_{y\in d_{i_{k-mean}}} R(\mathrm{x,y})\\ \underline{R_\vartheta}_{\,k-mean}d_i(\mathrm{x}) &= \min_{y\notin d_{i_{k-mean}}} \sqrt{1-R^2(\mathrm{x,y})}\\ \overline{R_\sigma}_{\,k-mean}d_i(\mathrm{x}) &= \max_{y\in d_{i_{k-mean}}} 1-\sqrt{1-R^2(\mathrm{x,y})}\end{aligned}\tag{11}$$

$$\underline{R_S}_{\,k-median}d_i(\mathrm{x}) = \min_{y\notin d_{k-median}} 1 - R(\mathrm{x,y})$$

$$\overline{R_{T}}_{\,k-median} d_i(\mathrm{x}) = \max_{y \in d_{i_{k-median}}} R(\mathrm{x},\mathrm{y})$$
$$\underline{R_{\vartheta}}_{\,k-median} d_i(\mathrm{x}) = \min_{y \notin d_{i_{k-median}}} \sqrt{1-R^2(\mathrm{x},\mathrm{y})}$$
$$\overline{R_{\sigma}}_{\,k-median} d_i(\mathrm{x}) = \max_{y \in d_{i_{k-median}}} 1-\sqrt{1-R^2(\mathrm{x},\mathrm{y})} \qquad (12)$$

The aforementioned models do not compute the lower and upper approximations with respect to the nearest samples as they might be outliers. These new models use *k*-trimmed or the mean or the median of *k* nearest samples to compute the membership of fuzzy approximations. This way, the variation of approximations caused by outliers is expected to be reduced; thus, the new models may be robust.

Given a binary classification task, $x \in d_1$ is a normal sample, and $\mathrm{y}_1 \in d_2$ is an outlier close to *x* such that $R(x, y_1)$ = 0.9.While as a normal sample, $\mathrm{y}_2 \in d_2$ is the second nearest sample of *x* from d_2, and $R(x, y_2)$ = 0.2. As per the classical fuzzy rough set model, $\underline{R_S} d_1(\mathrm{x}) = 1 - R(x, y_1)$ = 1 − 0.9 = 0.1.However, if we use the 1-trimmed model, $\underline{R_S}_{1-trimmed} d_1(\mathrm{x}) = 1 - R(x, y_2)$ = 0.8. This way, the noisy sample is ignored in the new model. At the same time, assume $x_1 \in d_1$ is the second nearest sample of y_1 , and $R(x_1, y_1)$ = 0.88. According to the classical model, $\underline{R_S} d_2(\mathrm{y}_1) = 1 - R(x, y_1)$ = 1 − 0.9 =0.1 and as per the 1-trimmed model, $\underline{R_S}_{1-trimmed} d_2(\mathrm{x}) = 1 - R(x_1, y_1)$ = 0.12.

It is seen that although the nearest sample *x* is ignored, y_1 still obtains a small value of membership. In fact, the membership should be small enough since y_1 is a noisy sample. This example shows that the this model can not only reduce the influence of noisy samples on computation of approximations of normal samples but can recognize the noisy samples and give small memberships to them as well.

IV. RNN-FRS BASED FEATURE SELECTION(RNN-FS)

The Given a decision table $\langle U, A, D \rangle$ is a nonempty universe, A is the set of attributes and D is the decision attribute. $\forall B \in A$, the membership of an object $x \in U$ belonging to the positive region of D on B is defined as

$$POS_B^*(\mathrm{D})(\mathrm{x}) = \sup_{x \in U/D} \underline{B^*}(\mathrm{X})(\mathrm{x}) \qquad (13)$$

The RNN fuzzy dependency of decision D on B is defined as

$$\gamma_B(\mathrm{D}) = \frac{\sum_{x \in U} POS_B^*(\mathrm{D})(\mathrm{x})}{|U|} \qquad (14)$$

RNN fuzzy dependency (RNN-FD) can also be used to evaluate features. Previous section proved that RNN-FRS fuzzy lower approximation is robust to the mislabelled samples. We consider that RNN fuzzy dependency is also robust to the mislabelled samples in feature evaluation.

Based on the RNN fuzzy dependency a feature selection algorithm is designed, shown in Table 1. The algorithm employs RNN-FD as the feature evaluation function and the sequential forward selection as the search strategy. The output of the algorithm is a feature ranking $F' = \{\mathrm{f}_1', \mathrm{f}_2', ..., \mathrm{f}_{|F'|}'\}$ .Given the set F_{k-1}' with k-1 features selected, the k' feature is determined by With the ranking, we can get n feature subsets $F_1' = \{\mathrm{f}_1'\}$, $F_2' = \{\mathrm{f}_1', \mathrm{f}_2'\}, ..., F_{|F'|}' = \{\mathrm{f}_1', \mathrm{f}_2', ..., \mathrm{f}_{|F'|}'\}$.

Table 1: RNN-FS algorithm

Input	X,F	X is a sample set and F is a feature set
Output	F'	F' is a feature ranking

```
Begin
    Initialize      $F' = \phi$
        While       F$\neq \phi$
                    Find f= $\arg_f \max_{f \in F} \{\gamma_{F' \cup \{f\}}(\mathrm{D})\}$
                    $F' = F' \cup \{\mathrm{f}\}$ ,
                    F=F-{f}
        End
        Return          $F^{'}$
    End
```

Next, we use RNN-FRS classifier to cross-validate the classification accuracy of the data with these feature subsets. The feature subset with the highest classification accuracy is the final feature subset.

V. NUMERICAL EXPERIMENTS

The simulation process is carried on a machine having Intel(R) core (TM) 2 Duo processor 2.40 GHz and 2.00 GB of RAM. The MATLAB version used is R2012(a).The simulation was carried out with 4 data sets collected from University of California, Irvine(UCI)Machine Learning Repository[34].

A. Data sets

Four datasets from University of California, Irvine (UCI) Machine Learning Repository are used [34]. The information related to the datasets is shown in Table 2.

B. Dataset Split

In the process of Classification, the dataset is split into ten parts. The randomly chosen 90% of objects are used as the training set and the remainder 10% as the testing set.

Table 2: Summaries of data sets

Datasets	Samples	Feature	classes
Wine	178	13	3
Lung cancer	32	56	3
WPBC	198	30	2
Glass	214	9	6

C. Input Parameters

In RNN based feature selection Algorithm takes some parameter s as input.

1. data_array: It takes the UCI datasets mention above.
2. evaluator: k-mean based RNN-FRC classifier is used as evaluator.
3. delta= Gaussian kernel function parameter. Here it's value taken as 0.15.
4. k = represent number of nearest neighbour. It is used for k-trimmed, k-mean and k-median.

D. Simulation Results

Firstly, we select features with the algorithm in Table 1 with real-world dataset. In this algorithm we use lower approximation based RNN-FD for RNN-FRS. We know robust nearest neighbour fuzzy rough classifier (RNN-FRC) are best for classification purposes [35]. It is of three types and they are k-trimmed, k-mean, k-median based RNN-FRC .So we use RNN-FRC(here k-mean classifier is used) to cross-validate the classification accuracy of the data set with the feature subsets $F_m^{'}\left(\mathrm{m}=1,2,...,\left|F_1^{'}\right|\right)$ composed of the first m features in the ranking. The feature subset with the highest classification accuracy is the final feature subset. Then to find features using fuzzy rough feature selection (FR-FS), replace RNN-FD with FD and select features with the same method.

The number of feature is selected for different datasets by different feature selection algorithm is shown below.

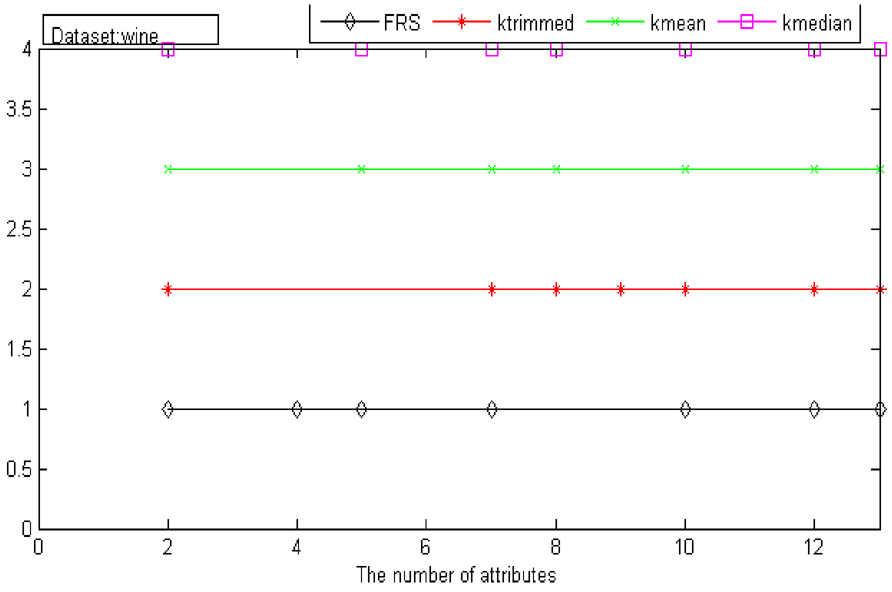


Fig. 2.Feature selected for Wine dataset

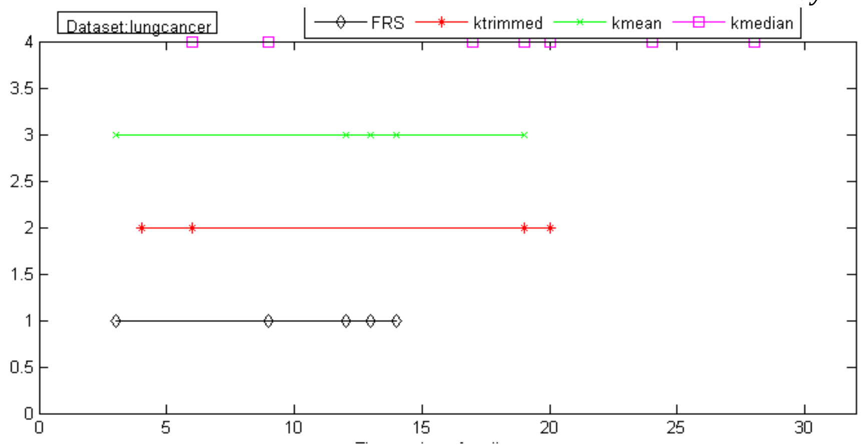


Fig. 3.Feature selected for Lung cancer dataset

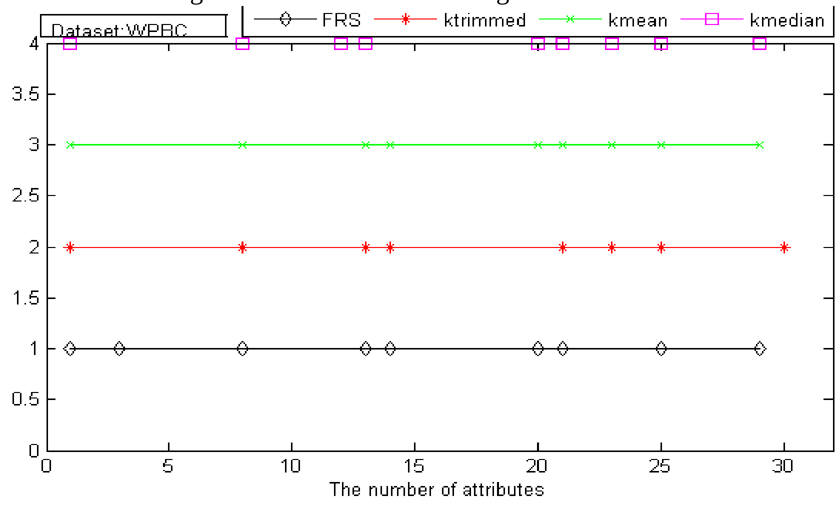
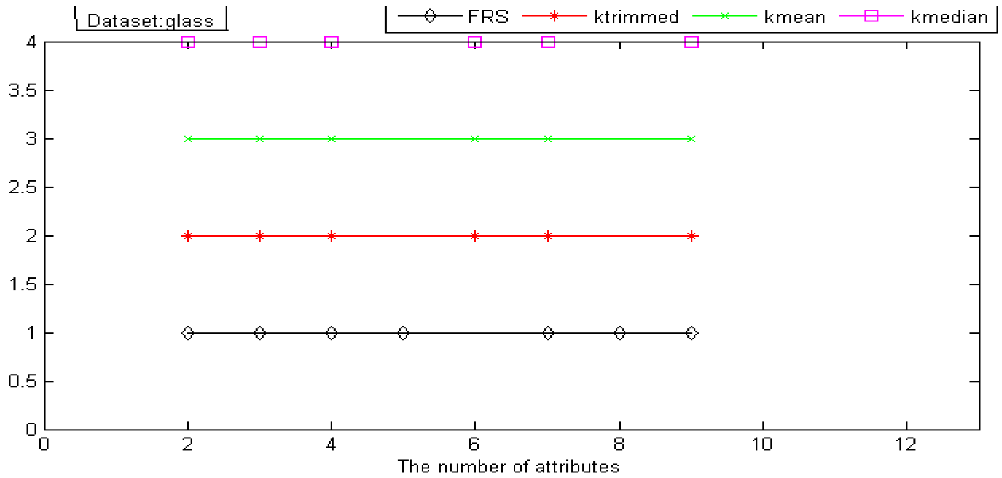


Fig. 4.Feature selected for WPBC dataset

Consider fig. 2, it describes feature selection for wine dataset, it has thirteen conditional features. For FRS based feature selection algorithm it selected seven feature sand they are 13,10,7,4,2,5,12.For k-trimmed based method it selected also seven feature and they are 13,10,7,2,8,12,9.For both k-mean and k-median based feature selection algorithm it selects seven features that are 13,107,25,8,12.

In fig. 3,using FRS based feature selection algorithm out of fifty six features of lung cancer, it selects five features, they are 9,3,12,14,13.Similarly, for k-trimmed, k-mean and k-median number of features selected are four, five and seven respectively and list of features are 19,6,4,20. For k-trimmed,19,12,13,3,14 for k-mean and 9,17,24,19,28,20,6 for k-median.

In WPBC, out of thirty features FRS selects nine features and they are1,25,13,8,14,20,21,3,29.K-trimmed selects eight features and they are 25,1,8,14,13,23,30,21.and for both k-mean and k-median nine features are selected and they are 1,25,13,8,14,20,21,23,29.But order of selected features are different. It is shown in Fig. 4.



.

Fig. 5.Feature selected for glass dataset

For glass dataset, FRS selects seven features out of nine and all other selects six features. For FRS features are 8,7,2,5,4,3,9 and for all other three features are 6, 3, 9, 4, 7, 2. It is shown in fig. 5.

Table 3: Number of feature selected and their classification accuracy on real world dataset

Datasets	**N**	**FRS**		**RNN-FRS**					
		n	**Acc(%)**	**k-trimmed**		**k-mean**		**k-median**	
				n	**Acc(%)**	**n**	**Acc(%)**	**n**	**Acc(%)**
Wine	13	7	97.19	7	98.31	7	97.63	7	97.50
Lung-cancer	32	5	56.25	4	71.88	5	62.50	7	65.63
WPBC	32	9	74.76	8	75.76	9	75.76	9	77.78
Glass	9	7	64.49	6	67.76	6	70.49	6	67.29
Average		7	73.17	6	78.42	7	76.49	7	77.05

The number of features selected and classification accuracies are shown in Table 3, where N is total number of feature, n is the number of features selected and Acc is classification accuracy. It is shown that, with RNN-FD as the feature evaluation, the feature subsets selected can produce higher classification accuracies.

In fig. 6, individual dataset and their classification accuracy evaluated by different feature selection methods are shown. The average number of feature is selected and their corresponding average classification is shown in figure 7.It shows all RNN-FRS based feature selection method produces higher classification accuracy than FRS based feature selection method (FR-FS). K-trimmed based feature selection method selects minimum features and provides highest classification accuracy among all RNN-FRS based feature selection methods (RNN-FS).

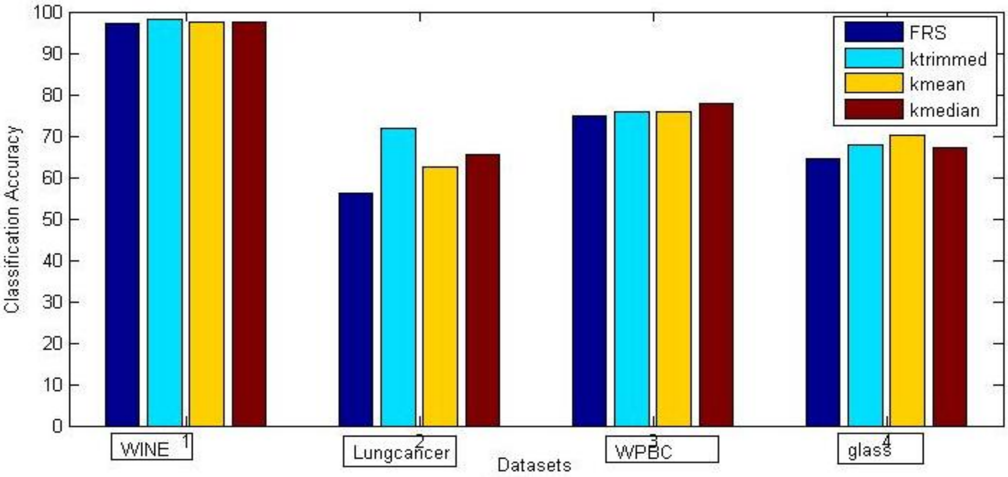


Figure 6.Classification accuracy of different datasets

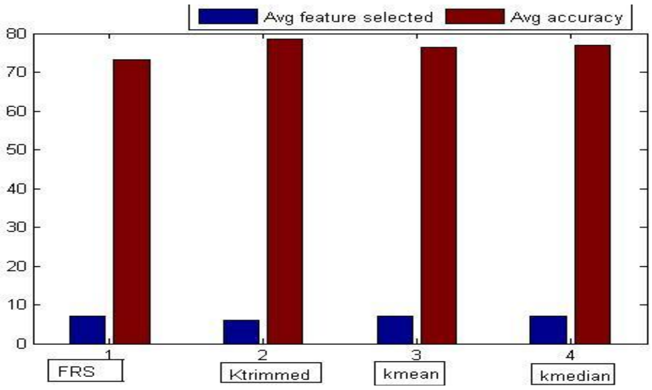


Figure 7.Average feature selected and classification accuracy of different methods

In this work, we use the classification accuracies of feature subsets to evaluate the robustness of measures. The higher the classification accuracy is, the stronger the robustness of the measure is. Therefore, all the algorithms of RNN-FS are more robust than FR-FS and k-trimmed is the best among them.

VI. CONCLUSIONS

Feature selection plays an important role in pattern classification systems. Feature evaluation function used to compute the quality of feature is a key issue in feature selection. In this chapter we successfully developed a algorithm for feature selection(RNN-FS).We also proved that the RNN-FS algorithm performs better to selecting important features and the dataset generated through the selected features provides good classification accuracy. The RNN based feature selection performs better than fuzzy rough set based feature selection and k-trimmed based feature selection method of RNN-FRS best among all.

.

REFERENCES

[1] Z. Pawlak, Rough sets, International Journal of Computer and Information Sciences 11 (1982) 341–356.

[2] Studies in Fuzziness and Soft Computing L. Polkowski, A. Skowron (Eds.), Rough Sets in Knowledge Discovery: Applications, Case Studies, and Software Systems, vol. 19, Physica-Verlag, Heidelberg, New York, 1998.D.

[3] R.W. Swiniarski, A. Skowron, Rough set methods in feature selection and recognition, Pattern Recognition Letters 24 (2003) 833–849

[4] Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, International Journal of General Systems 17 (1990) 191–209.

[5] J.-S. Mi, Y. Leung, H.-Y. Zhao, T. Feng, Generalized fuzzy rough sets determined by a triangular norm, Information Sciences 178 (2008) 3203–3213.

[6] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks 5 (1994) 531–549.

[7] A.B. David, H. Wang, A formalism for relevance and its application in feature subset selection, Machine Learning 41 (2000) 175–195.

[8] M.A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, in: Proceedings of the 17th International Conference on Machine Learning, 2000, pp. 359–366.

[9] Q.H. Hu, D.R. Yu, J.F. Liu, C.X. Wu, Neighborhood rough set based heterogeneous feature subset selection, Information Sciences 178 (2008) 3577–3594.

[10] N. Kwak, C.H. Choi, Input feature selection by mutual information based on Parzen window, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 1667–1671.

[11] H.C. Peng, F.H. 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 (2005) 1226–1238.

[12] R.W. Swiniarski, A. Skowron, Rough set methods in feature selection and recognition, Pattern Recognition Letters 24 (2003) 833–849.

[13] Q.H. Hu, D.R. Yu, J.F. Liu, C.X. Wu, Neighborhood rough set based heterogeneous feature subset selection, Information Sciences 178 (2008) 3577–3594.

[14] R. Jensen and Q. Shen, "Fuzzy-rough sets assisted attribute selection," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 1, pp. 73–89, Feb. 2007

[15] Q. Shen, R. Jensen, Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring, Pattern Recognition 37 (2004) 1351–1363.

[16] L. Yun, L.L. Bao, Feature selection based on loss-margin of nearest neighbor classification, Pattern Recognition 42 (2009) 1914–1921.

[17] G.-B. Ran, N. Amir, T. Naftali, Margin based feature selection-theory and algorithms, in: ACM International Conference Proceeding Series, Proceedings of the 21st International Conference on Machine Learning, vol. 69, 2004.

[18] Y.J. Sun, Iterative RELIEF for feature weighting: algorithms, theories, and applications, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2007) 1035–1051.

[19] H. Liu, H. Motoda (Eds.), Feature Selection for Knowledge Discovery and Data Mining, Kluwer Academic Publishers., Boston, 1998.

[20] L.J. Ke, Z.R. Feng, Z.G. Ren, An efficient ant colony optimization approach to attribute reduction in rough set theory, Pattern Recognition Letters 29 (2008) 1351–1357.

[21] P. Pudil, J. Novovicova, J. Kittler, Floating search methods in feature selection, Pattern Recognition Letters 15 (1994) 1119–1125.

[22] P. Somol, P. Pudil, J. Novoviova, P. Paclik, Adaptive floating search methods in feature selection, Pattern Recognition Letters 20 (1999) 1157–1163.

[23] H.C. Peng, F.H. 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 (2005) 1226–1238.

[24] Y.Y. Yao, S.K.M. Wong, P. Lingras, A decision-theoretic rough set model, in: Z.W. Ras, M. Zemankova, M.L. Emrich (Eds.), Methodologies for Intelligent Systems, vol. 5, New York, 1990, pp. 17–24.

[25] Y.Y. Yao, Y. Zhao, Attribute reduction in decision-theoretic rough set models, Information Sciences 178 (2008) 3356–3373.

[26] W. Ziarko, Variable precision rough set model, Journal of Computer and System Sciences 46 (1993) 39–59.

[27] Q.H. Hu, Z.X. Xie, D.Y. Yu, Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation, Pattern Recognition 40 (2007) 3509–3521.

[28] G.Y. Wang, J. Zhao, J.J. An, Y. Wu, A comparative study of algebra viewpoint and information viewpoint in attribute reduction, Fundamenta Informaticae 68 (2005) 289–301.

[29] C. Cornelis, M. De Cock, and A. M. Radzikowska, "Vaguely quantified rough sets," in *Proc. 11th Int. Conf. Rough Sets, Fuzzy Sets, Data Mining Granular Comput.*, 2007, pp. 87–94.

[30] C. Cornelis and R. Jensen, "A noise-tolerant approach to fuzzy-rough feature selection," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 2008, pp. 1598– 1605.

[31] S.Y. Zhao, E.C.C. Tsang, D.G. Chen, The model of fuzzy variable precision rough sets, IEEE Transactions on Fuzzy Systems 17 (2009) 451–467.

[32] Q. H. Hu, S. An, and D. R. Yu, "Soft fuzzy rough sets for robust feature evaluation and selection," *Inf. Sci.*, vol. 180, pp. 4384-4400, 2010.

[33] Qinghua Hu, Lei Zhang, Shuang An, David Zhang, Daren Yu. On robust fuzzy rough set models. IEEE Transactions on Fuzzy Systems. 2012, 20(4): 636-651.

[34] A. Asuncion and D. J. Newman (2007). UCI machine learning repository,School Inf. Comput. Sci., Univ. California, Irvine, [Online].Available:http://www.ics.uci.edu/mlearn/MLRepository.html
Bichitrananda Behera,S.Sabat ,Comparison of Robust Nearest Neighbor Fuzzy Rough Classifier with KNN and NEC classifiers, International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2056-2062.