A Novel Approach Feature Selection based on Neighborhood Positive Region (NPR)

L.H. Patil

Research Scholar, Department of Computer
Science and Engineering, Sant Gadge Baba
Amravati University, Amravati, India

Mohammed Atique, PhD

Associate Professor, Department of Computer
Science and Engineering, Sant Gadge Baba
Amravati University, Amravati, India

ABSTRACT
Due to increase in large number of document on the internet data mining becomes an important key parameter. Numerous data mining techniques are being carried for extracting the valuable information such as clustering, classification and cluster analysis. In the field of machine learning, pattern recognition and data mining, feature selection also called as attribute reduction becomes a challenging problem. Also the key lies in reducing the attributes and selecting the relevant features. Hence, to overcome the issues of attribute reduction we proposed Neighborhood positive region (NPR) based on rough set theory. In this paper we have shown the experimental result of NPR is implemented on three UCI data sets which show the computational time and reduced features.

Keywords
Feature selection, Neighborhood positive region, Classifier.

1. INTRODUCTION
Feature selection also called as attribute reduction has become an important steps in the field of data mining, pattern recognition and machine learning tasks. Feature selection based on rough set theory is also called attribute reduction. Recently, in rough set theory the datasets have grown drastically, both in number and dimensionality of items. Ten, hundreds and even thousands of attributes are available in the real time databases. [1][2]A key problem in rough set theory is executing attribute reduction for a data set. Rough set theory has been conceived as a powerful soft computing tool to analyze various types of data, and is also a specific framework of selecting useful features it becomes difficult to recognize the task of relevant attributes, which decline the performance of machine learning problem. Rough set theory [10] proposed by Pawlak, is an important mathematical tool, which can be used for handling data sets with imprecision and uncertain information. It can be employed to study attribute reduction and feature selection in information systems. The main application of rough set theory is to obtain a concept approximation of a universe by two definable subsets called lower and upper approximations. It has been known that lower and upper approximation operators in Pawlak's rough set are defined by an equivalence relation [12]. [6][7] Developed a common accelerator to improve the time efficiency of a heuristic search process. According to the accelerator, certain objects are deleted from the universe every time when a new attribute is selected and added into the core. Rough Set Attribute Reduction (RSAR) [3] is the concept of indiscernibility. Rough Set Attribute Reduction (RSAR) provides a filter-based technique by which knowledge can be extracted from a domain in a concise way; retaining the information content even as reducing the amount of knowledge involved. A neighborhood relation is used to generate a family of granules from the universe characterized with numerical features. These neighborhood granules are used to approximate the decision classes. Considering this idea, a neighborhood rough set model was been constructed [4][6][7]. Till now no work has been deal with heterogeneous features with tolerance rough set model yet. In reality the neighborhood rough set model is expected as a natural generalization of Pawlak's rough set model. Hence, it is therefore important to reduce dimensionality of the data to smaller set of features and relevant information for decreasing the cost in storing and reduction in the processing time. To overcome the issue of feature selection and attribute reduction, few attributes can be omitted, which will not seriously affect on classification accuracy.

The two key issues in rough set methodology are granulation and approximation.

Granulation segments the space data into different subsets on certain criterion. The generated subsets are called elemental granules or elemental concepts. Approximation, approximately describe arbitrary subset of the space data with these elemental concepts.

The main contribution of the work lies on three fold. 1) It extends the Neighborhood Positive rough set model to deal with features and consistency, neighborhood sizes and attributes 2). Computing time and finding the relevant features. 3) Finding the accuracy using KNN and NB classifier

2. PRELIMINARIES OF ROUGH SET
Rough set theory has been identified as the one of the most powerful tool in the field of data mining for feature selection. Feature selection is frequently used in data mining and machine learning task. The main objective of rough set to feature selection is to find the minimal subset from the original features. It also predicts the decision concept of the original features. The criteria for selecting feature are mainly determined by the usefulness of the feature. This usefulness of a feature is determined by its redundancy and relevant. [10] has proposed the rough set theory for knowledge discovery in databases. It is based on the basic concept of an upper and lower approximation of a set, the approximation space and set of model. The basic mechanism of rough set model is described in figure 4.1
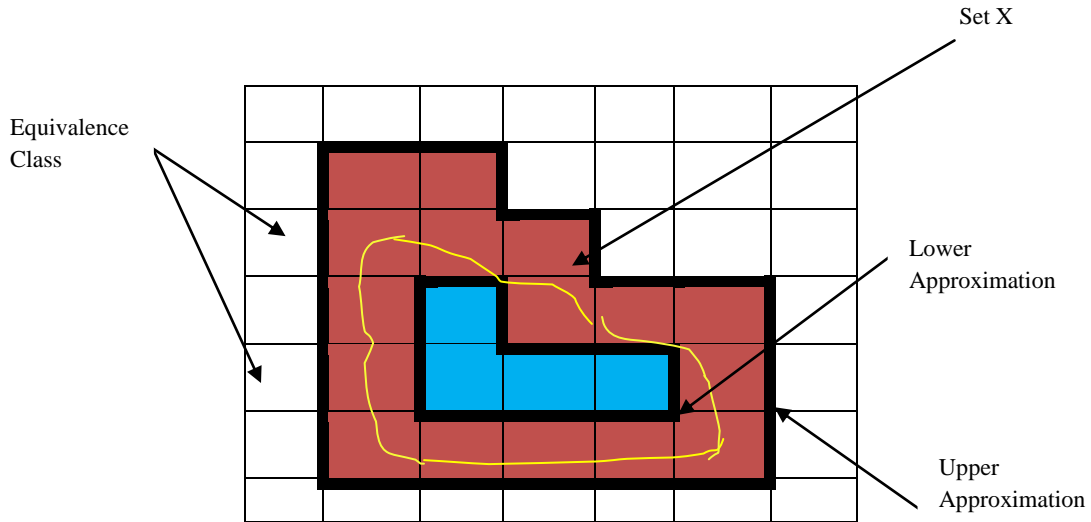
Figure 1: Basic mechanism of Rough set

An information system is defined as

$S = <U, Q, V, f>$ where U is the universe, a finite set of N objects($x_1, x_2, ..., x_n$) a nonempty set. Q is a finite set of attributes. $V = \cup_{q \in Q} V_q$ (Where V_q is a domain of the attribute q), f: U X Q→V is called information function such that $f(x, q) \in V_q$ for every q$\in$Q, x$\in$U. A subset of attributes $A \sqsubseteq$ Q defines an indiscrenibilty relation on U.

$IND(A) = \{ (x, y) \in \cup : \text{for all } a \in A, f(x, a) = f(y, a)\}$

denoted also $\overline{A}$. The same can be defined as a decision table. $DT = <U, C \cup D, V, f>$ Where C is a set of condition attribute, D is the set of Decision attribute. $V = \cup_{q \in C \cup D} V_q$ Where V_q is the set of domain of an attribute q $\in$ Q, f: UX(C $\cup$ D) → V is a total decision function such that $f(x, q) \in V_q$ for every q $\in$ Q and x $\in$ V.

For the given S a subset of attribute $A \subseteq Q$ determines the approximation space. $AS = (U, IND(A))$ in S. For given $A \subseteq Q$ and $X \subseteq U$, the A-lower approximation $\underline{A}X$ of the set X in AS and the A-upper approximation $\overline{A}X$ of the set X in AS are defined as follows:

$$\underline{A}X = \{ x \in U : [x]_A \subseteq X \} = \cup \{ Y \in A^* : Y \subseteq X \}$$

$$\overline{A}X = \{ x \in U : [x]_A \cap X \neq \emptyset \} = \cup \{ Y \in A^* : Y \cap X \neq \emptyset \}$$

Rough set provides the concept to determines for a given information system the most important attributes. The main idea of the reduct is fundamental for rough set theory. An essential part of an information system is a reduct which is related to a subset of attributes. Another important part is a core. The reduct and core is an important concept of rough set theory which is generally used for feature selection and attribute reduction. Rough set theory determines the significance measures, degree of attributes and dependency. For the given information system $S = <U, Q, V, f>$ with the condition and decision attribute. $Q = C \cup D$ $A \subset C$ can be defined as A positive region $POS_A(D)$ in the relation $IND(D)$ as

$$POS_A(D) = \cup \{ \underline{A}X : X \in IND(D) \}$$

$POS_A(D)$ contains all the objects in U. A positive region for nay two subsets of attributes $A, B \in Q$ in the information system S. The subset of attributes $B \in Q$ defines the indiscrenibility relation $IND(B)$ which defines the classification $B * (U/IND(B))$ with respect to subset A. Positive region of B is defined as

$$POS_A(B) = \bigcup_{X \in B} \underline{A} X$$

Positive region of B contains all the object. The cardinality of positive region B defines a measure $Y_A(B)$ of dependency of the set of attributes B on A

$$Y_A(B) = \frac{Card(POS_A(B))}{Card(U)}$$

From the information system S a set of all attributes B depends on A in S, which is denoted as A→ B; iff satisfies the equivalence relation $IND(A) \subseteq IND(B)$. Two sets A and B are independent of S if neither A→B nor B→A hold. The dependency of set B to degree K to the set A in S is denoted as

$$A \xrightarrow{k} B, \qquad 0 \leq k \leq 1 \quad \text{if } k = \gamma_A(B)$$

Definition: Rough set defines a measure of significance or coefficient of significance of the attribute a $\in$ A from set A with respect to classification $B * (U/IND(B))$ generated by set B.

$$\mu_{A,B}(a) = \frac{card(POS_A(B)) - Card(POS_{A-\{a\}}(B))}{Card\ U}$$

A significance of attribute a in the set $A \subset Q$ can be computed with respect to original classification Q*.

3. ENTROPY BASED REDUCTION
A filter-based feature selection is entropy-based reduction (EBR). This approach is based on the entropy heuristic employed by machine learning techniques such as C4.5. [1] EBR is concerned with examining a dataset and determining those attributes that provide the most gain in information. The entropy of attribute *A* with respect to the conclusion *C* is defined as:

$$H(C \mid A) = -\sum_{j=1}^{m} P(C_i \mid a_i) \log_2 P(C_i \mid a_i)$$

EBR(C,D)
C, the set of all conditional features;
D, the set of decision features.

(1) $R \leftarrow \{\}$
(2) do
(3) $T \leftarrow R$
(4) $\forall x \in (C - R)$
(5) (5) if $H(R \cup \{x\}) < H(T)$
(6) $T \leftarrow R \cup \{x\}$
(7) $R \leftarrow T$
(8) until $H(D \mid R) = H(D \mid C)$
(9) return R

Considering the entropy measure, the algorithm can be modified by using rough set based attribute reduction. This works on dealing with subset of attributes instead on individual attributes. EBR algorithm requires no threshold. It searches the attributes when the best feature subset is stopped as the resulting subset entropy is equal to that of the entire feature set. For consistent data, the final entropy of the subset becomes zero. Any subset with entropy of 0 will have a corresponding Rough set dependency of 1. Hence it is used for finding rough set reduct if the data is consistent.

4. QUICKREDUCT ROUGH SET ATTRIBUTE REDUCTION

Let $I = (U, A)$ be an information system, where U is non-empty set of finite objects and A is a non empty finite set of attributes such that a: $U \rightarrow V_a$ for every $\in A$, V_a is the set of values that attribute a may take. With any $P \subseteq A$ these is an associated equivalence relation IND(P)

$$IND(P) = \{(x, y) \in U^2 \mid \forall a \in P, a(x) = a(y)\}$$

The Partitioning of U generated by IND(P) is denoted U/IND(P) then, x& y are indiscernible of attribute from-P. the P-indiscernibility relation denoted as $[x]_p$.

Let $X \subseteq U$, X can be approximated within P by constructing P - lower and P – upper approximation of X.

$$\underline{P}X = \{x \mid [x]_p \subseteq X\}$$

$$\overline{P}X = \{x \mid [x]_p \cap X \neq \emptyset\}$$

Let P and Q be equivalence relation over U and positive region be defined as

$$POS_p(Q) = \sum_{X \in U/Q} \underline{P}X$$

[1]The positive region contains all objects of U which is classified as U/Q using information in attribute P. considering the definition of positive region, the rough set degree of dependency of a set of attributes Q on set of attributes P be defined as, For $P, Q \subset A$ it is said that Q depends on P in a degree k $(0 \le k \le 1)$ denoted $P \rightarrow_k Q$ if

$$k = \gamma_p(Q) = \frac{\mid POS_p(Q) \mid}{\mid U \mid}$$

The reduction of attributes can be achieved by comparing equivalence relation from the set of attributes. A reduct R is defined as a subset of minimal cardinality of conditional attribute set C such that $\gamma_R(D) = \gamma_C(D)$.

The QUICKREDUCT Algorithm is shown in fig 4.

QUICKREDUCT (C,D)
C, the set of all conditional features
D, the set of all decision features.

(1) $R \leftarrow \{\}$
(2) do
(3) $T \leftarrow R$
(4) $\forall x \in (C - R)$
    a. If $\gamma_{R \cup \{x\}}(D) > \gamma_r(D)$
(5) $T \leftarrow R \cup \{x\}$
(6) $R \leftarrow T$
(7) until $\gamma_r(D) = \gamma_C(D)$
(8) return R

5. VARIABLE PRECISION ROUGH SET

[11] has proposed VPRS and extends Rough set theory by the relaxation of subset operator. It analyses and identifies data patterns as a Statistical trends rather than functional. The main objective of VPRS is to allow objects to classify with an smaller error than a predefined level. This means it introduces a threshold relaxes the Rough set notion of requiring no information outside the dataset. The relative classification error is defined as for $X, Y \subseteq U$:

$$c(X, Y) = 1 - \frac{\mid X \cap Y \mid}{\mid X \mid}$$

By observing that $c(X,Y) = 0$ if and only if $X \subseteq Y$. A degree of inclusion can be defined at a certain level of error β in classification

$$X \subseteq_\beta Y \text{ iff } c(X, Y) \le \beta, 0 \le \beta < 0.5$$

The β-upper and β-lower approximation of a set X can be defined as

$$\underline{P}_\beta X = \cup \{[X]_P \in U/_P \mid [x]_P \subseteq_\beta X\}$$

$$\overline{P_\beta} X = \cup \{[X]_P \in U/_P \mid c([x]_p, X) < 1 - \beta\}$$

where $\overline{P_\beta} X = \underline{P}X$ for $\beta = 0$.

Thus the positive, negative and boundary region from the original rough set theory can be written as:

$$POS_{P,\beta}(Q) = \bigcup_{X \in \frac{U}{Q}} \underline{P}_\beta X$$

$$NEG_{P,\beta}(Q) = U - \bigcup_{X \in \frac{U}{Q}} \overline{P_\beta} X$$

$$BND_{P,\beta} = (Q) = \bigcup_{X \in \frac{U}{Q}} \overline{P}_\beta X - \bigcup_{X \in \frac{U}{Q}} \underline{P}_\beta X$$

where P is an Equivalent relation on U. This calculates the dependency and thus determines the $\beta - reducts$. The dependency function is defined as:

$$\gamma_{P,\beta}(Q) = \frac{\mid POS_{P,\beta}(Q) \mid}{\mid U \mid}$$

In this QUICKREDUCT algorithm can be adapted which incorporates the reduction method built by VPRS. Considering the suitable values β- lower approximation, $\beta -$ Positive region and β- dependency replaces the calculation. This result in more approximate final reducts for better generalization.

6. PROPOSED NEIGHBORHOOD ROUGH SET

In neighbourhood rough set model the basic concepts of rough set are reviewed. We suppose that the universe data used is denoted by information system IS $=< U, A >$, where U is non empty and finite set of samples $\{s_1, s_2, \dots, s_n\}$ called a universe. A is a set of attributes $\{a_1, a_2, \dots, a_n\}$ to characterize the samples $< U, A >$ is called as decision table, if $A = C \cup D$ where C is condition attribute and D is a decision attribute. Given an arbitrary variable $s_i \in U$ and $B \subseteq C$ the neighbourhood $\delta_B(s_i) = \{s_j | s_j \in U, \Delta^B(s_i, s_j) \leq \delta\}$ where Δ is a distance function for $\forall_{s_1, s_2, s_3} \in U$, it usually satisfies,

1. $\Delta(s_1, s_2) \geq 0, \Delta(s_1, s_2) = 0$ if and only if.

2. $\Delta(s_1, s_2) = \Delta(s_1, s_2);$

3. $\Delta(s_1, s_3) \leq \Delta(s_1, s_2) + \Delta(s_1, s_3);$

These three metric distance function are mostly used in machine learning, data mining and pattern recognition task. A general metric named minkowsky distance is defined as: consider two objects s_1, s_2 in N- dimensional space $A = \{a_1, a_2, \dots, a_n\}$, $f(s, a_i)$ denotes the value of sample s in the ith attribute a_i. Therefore $\Delta_P(s_1, s_2) = (\sum_{i=1}^{N} |f(s_1, a_i) - f(s_2, a_i)|^P)^{1/p}$ where

If i) $p = 1$ called as manhattan distance $\Delta_1 = 1$ ii) if $p = 2$ called as Euclidean distance $\Delta_1 = 2$ iii) if $p = \infty$ called as chebychev distance for $\Delta_1 = \infty$

$\delta_B(s_i)$ is the neighbourhood center with sample s_i and the size of neighbourhood depends on the threshold δ. If δ has greater value than more samples get connected into neighbourhood of s_i.

In Neighbourhood Rough set model the two key factors play an important role i) used distance and ii) threshold δ.

The used distance determines the shape of neighbourhood. The threshold δ controls the size of neighbourhood. If the $\delta = 0$ then the neighbourhood degrades to an equivalent class and the samples of neighbourhood rough set models degenerates to one. Hence the neighbourhood rough set model is the natural generalization of the rough set. To deal with different attributes the different definitions are used to compute neighbourhood samples $(s_0, s_1, \dots, s_n)$ with different attributes. Let $B_1 \subseteq A$ and $B_2 \subseteq A$ be the different attributes, so the neighbourhood granules for sample *s* for B_1, B_2 and $B_1 \cup B_2$ are

1. $\delta_{B_1}(s) = \{s_i | \Delta_{B_1}(s, s_i) \leq \delta, s_i \in U\};$

2. $\delta_{B_2}(s) = \{s_i | \Delta_{B_2}(s, s_i) = 0, s_i \in U\};$

3. $\delta_{B_1 \cup B_2}(s) = \{s_i | \Delta_{B_1}(s, s_i) \leq \delta^\wedge \Delta_{B_2}(s, s_i) = 0, s_i \in U\};$

Where $\wedge$ denotes "and" operator.

Where 1) is defined as numerical attribute. 2) is defined for categorical attribute. 3) mixed categorical and numerical attribute. According to this definition, the samples in a

neighborhood granule have the similar values in terms of categorical attributes and the distance in term of numerical attributes is less than threshold δ. Besides these there are many distance function for numerical and categorical attributes such as HVDM, Heterogeneous Euclidean-Overlap Metric function (HEOM), Interpolated Value Difference Metric IVDM, Value Difference Metric (VDM), and Heterogeneous Value Difference Metric (HVDM). HVDM is defined as

$$HVDM(x, y) = \sqrt{\sum_{i=1}^{m} w_{a_i} \times d_{a_i}^2(x_{a_i}, y_{a_i})}$$

Where $d_{a_i}(x_{a_i}, y_{a_i}) =$

$$\begin{cases} 1, \text{if x or y is unknown} \\ normalized\_vdm_a(x, y), \text{if a is nominal} \\ normalized\_diff_a(x, y), \text{if a is linear} \end{cases}$$

$$normalized_{diff_a}(x, y) = \frac{|x - y|}{4\sigma_a}$$

Where σ_a is the standard deviation of the numeric values of attributes a.

$$N1: normalized\_vdm1_a(x, y) = \sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|$$

$$N2: normalized\_vdm2_a(x, y) = \sqrt{\sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^2}$$

$$N3: normalized\_vdm3_a(x, y) = \sqrt{C * \sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^2}$$

If a set of object and neighbourhood relation N over U is called as neighbourhood approximation space. For any $S \subseteq C$, two objects called upper and lower approximation of S in $< U, N >$ is defined as

$$\underline{N}S = \{s_i | \delta(s_i) \subseteq S, s_i \in U\}$$

$$\overline{N}S = \{s_i | \delta(s_i) \cap S \neq \emptyset, s_i \in U\}$$

Obviously $\underline{N}S \subseteq S \subseteq \overline{N}S$. The boundary region of S in the approximation space is defined as $BNS = \overline{N}S - \underline{N}S$. The size of the boundary effects on the degree of roughness of S in the approximation space $< U, N >$. The size of the boundary region depends on attribute X to hold U and threshold δ.

6.1 Feature Selection based on Neighborhood Positive Region (NPR)

Pawlak proposed the concept of positive region, which measure the significance of a condition attribute from a decision table. In this a neighborhood positive region has been proposed for generating a neighborhood relation on the universe.

Definition 1. Let the decision table IS $=< U, C \cup D, N >$ $S_1, S_2, \dots, S_N$ are the object subsets with decisions 1 to N; $\delta_B(s_i)$ is the neighborhood information granule generated by attribute $B \subseteq C$, the lower and upper approximation of decision D with respect to attributes B are defined as

$$\underline{N_B}D = \cup_{i=1}^{N} \underline{N_B}S_i, \quad \overline{N_B}D = \cup_{i=1}^{N} \overline{N_B}S_i,$$

Where

$\underline{N_B}S = \{s_i| \quad \delta_B(s_i) \subseteq S, s_i \in U\}, \overline{N_B}S = \{s_i| \quad \delta_B(s_i) \cap S \neq \emptyset, s_i \in U\}$

The decision boundary region of D with respect to attributes B is defined as

$BN(D) = \overline{N_B}D - \underline{N_B}D$

The lower approximation of the decision is defined as the union of the lower approximation of each decision class. The lower approximation of the decision is also called the neighborhood positive region of the decision, denoted by $POS_B(D)$. $POS_B(D)$ is the subset of objects whose neighborhood granules consistently belong to one the decision classes.

Definition 2: Given the decision table $IS = < U, C \cup D >$, the distance function Δ and neighborhood size δ, the dependency degree of D to NP_i is defined as $\gamma_{NP_i}(D) = \frac{\left|POS_{NP_i}^U(D)\right|}{|U|}$

Where $|\cdot|$ is the cardinality of a set. $\gamma_{NP}(D)$ is the ability of NP to approximate D. As $POS_{NP_i}(D) \subseteq U$, we have $0 \leq \gamma_{NP}(D) \leq 1$. we say D completely depends on NP and the decision system is consistent in terms of and δ. If $\gamma_{NP}(D) = 1$; otherwise, it can be D depends on NP in the degree of γ.

Dependency function totally depends on the size of the region between classes. Dependency function returns the approximation power of condition attribute. It is used to measure the significance of the attribute. The main objective of attribute selection is to search feature subset of attributes. In these the classification problem has the maximal consistency in the selected feature spaces. In rough set theory an attribute reduction is used to find the attribute subset that have the minimal attributes and retain the power of original features.

To design an algorithm three important parameters has to consider which significance measures, searching strategy and termination criteria. To improve the searching strategy of forward greedy search algorithm, the concept of neighborhood positive region has been introduced. In forward greedy search algorithm, we start the attribute with maxima inner significance and then assign the attribute with maxima outer significance into the subset in each loop, until it satisfies the stooping criteria and at the end we produce an attribute reduct.

Given decision table IS = $< U, C \cup D >$, condition partition can be obtain as $U/_C = \{X_1, X_2, ..., X_m\}$ and the decision partition as $U/_D = \{Y_1, Y_2, ..., Y_n\}$. Considering these notations attribute significance measures are defined as follow: Significance measures are computed for attribute evaluation function and then represents greedy search feature selection algorithm based on NPR.

Definition 3: Let the decision table IS = $< U, C \cup D >, B \subseteq C$ and $\forall a \in B$. The significance measure of a in B is defined as

$SIG_1^{inner}(a, B, D) = \gamma_B(D) - \gamma_{B-a}(D)$

where $\gamma_B(D) = \frac{|POS_B(D)|}{|U|}$

**Definition 4:** Let the decision table $IS = < U, C \cup D >, B \subseteq C$ and $\forall a \in C - B$. The significance measure of a in B is defined as

$SIG_1^{outer}(a, B, D) = \gamma_{B \cup a}(D) - \gamma_B(D)$.

The definitions are used to select an attribute reduction algorithm. From given decision table the intersection of all attribute reducts is said to be indispensible and is also called as core. Each core attribute must also be in the reduct attribute of decision table. The core may be an empty set.

Definition 5: Let the decision table IS =$< U, C \cup D >, B \subseteq A, \forall a \in B$, we say a is superfluous in B if $\gamma_{B-a}(D) = \gamma_B(D)$; otherwise, we say a is indispensable. We say attribute B is indispensable. We say attribute B is independent relative to the decision D if $\forall a \in B$ is indispensable.

The process of forward greedy search algorithm based on NPR is shown in Figure 2.
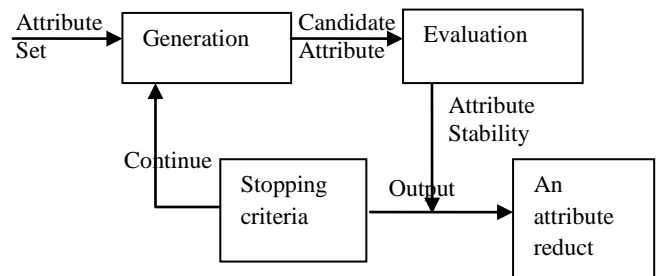
Figure 2: Forward greedy attribute reduction algorithm

The general forward greedy attribute reduction algorithm is written as follow:

Algorithm 1: A general forward greedy attribute reduction algorithm based on NPR

Input   1: Decision Table IS =$< U, C \cup D >$
        2. Delta δ // δ is the threshold to control the size of neighborhood.
Output: One reduct red
Step 1: red $\leftarrow \emptyset$ // red is the pool of selected attributes
Step 2: For each $a_k \in C -$ red
Step 3:   Compute $\gamma_{red \cup a_k}(D) = \frac{|POS_{B \cup a_k}(D)|}{|U|}$
Step 4:   Compute$SIG^{inner}(a_k. red, D) = \gamma_{red \cup a_k}(D) - \gamma_{red}(D)$
Step 5:   End.
Step 6: Compute $Sig^{inner}(a_k, C, D, U); k \leq |C|$
//$Sig^{inner}(a_k, C, D, U)$is the inner importance measure of attribute a_k
Step 7: Put a_k into red, where $Sig^{inner}(a_k, C, D, U) > 0$;
Step 8: While $EF(red, D) \neq EF(C, D)$ do// provides stopping criteria
     {
       red $\leftarrow$ red $\cup \{a_0\}$
      where
$Sig^{outer}(a_0, red, D) = max\{ Sig^{outer}(a_k, red, D), a_k \in C -$ red
     }//$Sig^{outer}(a_k, C, D)$is the outer importance of the attribute a_k
Step 9: Return red
Step 10: End

7. EXPERIMENTAL ANALYSIS

In this we evaluate the performance of general forward greedy attribute reduction algorithm and the proposed improved feature selection based on neighborhood positive approximation rough set algorithm. In this chapter three different methods are proposed and are compare the number of feature selected, computational time and classification accuracies of the general algorithm with proposed algorithm. The experimental results shows the features at the stage of documents preprocessing, the features subset selects with the general and proposed algorithm. The objective of experimental analysis is to compute the time efficiency of the proposed feature selection algorithm. The data used in the experiments are outlined in Table 1. Where the three datasets are used and downloaded from UCI machine learning databases. Three datasets are used to preprocess the data, to compare the computational time for general algorithm and the proposed algorithm. These algorithms are run on Personal Computer with Windows XP and Intel® Core™ i7 CPU 2.66 GHz, 4.00 GB memory. The software being used is MATLAB R2010b. In these the last two columns shows the features and classes used for further preprocessing. These features are widely used for feature extraction technique.

Table 1: Data set description

Sr.No	Data Set	Documents	Features	Classes
01	Reuters 21578	21	6539	04
02	Classic 4	54	1625	06
03	Newsgroup 20	52	1454	04

An experimental result shows that the number of features selected for NPR. Hence the stability of Improved NPR algorithm retains same the predictive power of original features. The computational time for Improved NPR largely reduces as compare to other algorithm. The details of computational time and number of feature selection for different threshold values are shown in Table 2-4. Threshold delta plays an important role in neighborhood positive approximation rough sets. It is considered as a parameter to control the size of the neighborhood. The computational time tested for two algorithms on three datasets Reuters 21578, Classic 04 and Newsgroup 20. Figure 4-6 shows the computational time for NPR with three different threshold values. Each one present computational time of NPR for delta values 0.01, 0.015 and 0.001. In these the computational time with NPR significantly reduces with same number of features. Other improved method EBR, QUICKREDUCT and VPRS are used and shown that how the feature and computational time is reduced. Table 5 shows the number of features and computational time for EBR, QUICKREDUCT and VPRS.

Table 2: Time and feature selection of the algorithms NPR with $\delta = 0.01$

Sr.No	Data Set	Features	NPR Features	NPR Time (s)
01	Reuters 21578	5677	448	60.00
02	Classic 4	1411	92	30.00
03	Newsgroup 20	976	71	20.00

**Table 3: Time and feature selection of the algorithms NPR with $\delta = 0.015$**

Sr.No	Data Set	Features	NPR Features	NPR Time (s)
01	Reuters 21578	5677	441	50.00
02	Classic 4	1411	86	25.00
03	Newsgroup 20	976	65	15.00

Table 4: Time and feature selection of the algorithms NPR with $\delta = 0.001$

Sr.No	Data Set	Features	NPR Features	NPR Time (s)
01	Reuters 21578	5677	465	74.00
02	Classic 4	1411	101	40.00
03	Newsgroup 20	976	76	25.00

Table 5: Time and feature selection of the algorithms EBR, QUICKREDUCT, VPRS ($\beta = 0.001$),NPR

SN	Data Set	Features	EBR Features	EBR Time (s)	QUICKREDUCT Features	QUICKREDUCT Time (s)	VPRS Features	VPRS Time (s)	NPR Features	NPR Time (s)
1	Reuters 21578	5677	451	55	461	65	461	65	448	60
2	Classic 4	1411	96	45	102	40	102	40	92	30
3	Newsgroup 20	976	78	25	79	25	79	25	71	20
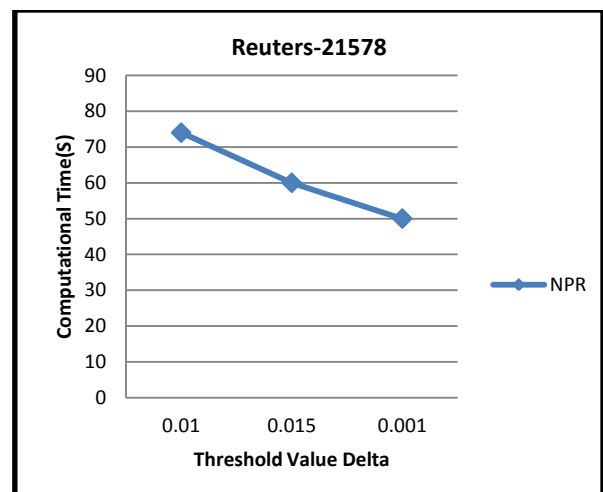
Figure 4: Comparison of computational time on Reuters 21578
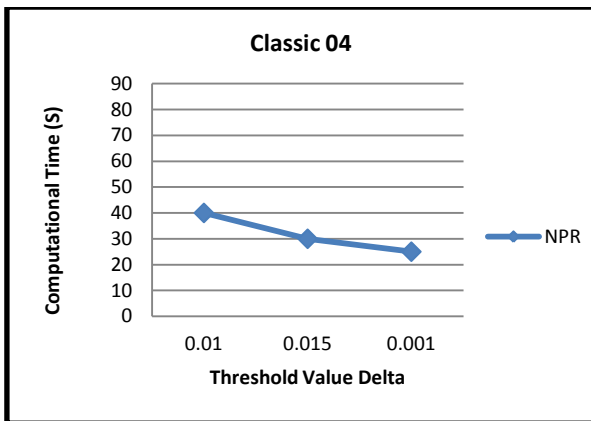
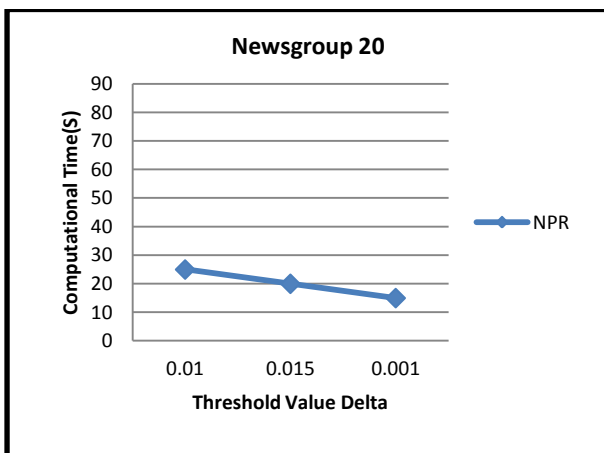**Figure 5 : Comparison of computational time on Classic 04**

Figure 6: Comparison of computational time on Newsgroup 20

For result validation of NPR algorithm we use NB and KNN classifier. This Classifier shows the accuracy of feature selection algorithm. The accuracy for NPR algorithm retains the same which shows that there is no loss of relevant information even after the modification in algorithm. The accuracy result for three different UCI data sets and feature selection with NB and KNN classifier is shown in table 6.

Table 6: The accuracy of feature selection algorithm with NB and KNN

Sr. No	Data set	Features	NPR	
			NB	KNN
1	Reuters 21578	465	97.85	98.85
2	Classic 4	101	96.14	97.36
3	Newsgroup 20	76	92.45	94.75

8. CONCLUSION

To overcome the limitation of existing feature selection algorithm, a new novel approach which is called as neighborhood positive region based on rough set have been proposed. This can be used to reduce computational time for feature selection in dealing with large data sets. Three different feature selection algorithm EBR, QUICKREDUCT and VPRS have been used for several representative algorithms which been modified. Each of the algorithms can choose the same feature subset as the original one. Experimental analysis on three UCI data sets shows that the proposed algorithm reduces the computational time of feature selection without losing its stability. The results show that the improved feature selection based on rough set model is more efficient concern with the stability and computational time in dealing with large datasets. In future the rough set model can implemented by using chi square statistic to find the probability of each lower approximation and final base clustering can be evaluated.

9. REFERENCES

[1] C Velayutham, K Thangavel,*"* Unsupervised quick reduct algorithm using rough set theory", Journal of Electronic Science, 2011

[2] K. Kira, L.A. Rendell, "The feature selection problem: traditional methods and a new algorithm", in: Proceedings of AAAI-92, pp.129–134,1992.

[3] M. Modrzejewski, "Feature selection using rough set theory", in: Proceedings of European Conference on Machine Learning, pp. 213–226, 1993.

[4] R. Jensen, Q. Shen, "Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches", IEEE Transactions on Knowledge and Data Engineering 16 (12) 1457–1471, 2004.

[5] Qinghua Hu, Daren Yu, Jinfu Liu, Congxin Wu, "Neighborhood rough set based heterogeneous feature subset selection", Information Sciences 178, 3577–3594, 2008.

[6] W. Pedrycz, G. Vukovich, "Feature analysis through information granulation and fuzzy sets", Pattern Recognition 35 825–834, 2002

[7] Yuhua Qian, Jiye Liang, Witold Pedrycz, Chuangyin Dang, Positive approximation: An accelerator for attribute reduction in rough set theory, Artificial Intelligence 174, 597–618, 2010

[8] Yuhua Qiana, Jiye Lianga, Witold Pedrycz, Chuangyin Dang, "An efficient accelerator for attribute reduction from incomplete data in rough set framework", Recognition 44, 1658–1670. 2011

[9] C.S. Yang, L. Shu, "Attribute reduction algorithm of incomplete decision table based on tolerance relation", Computer Technology and Development 16 (9), 68–69 72, 2006

[10] R. W. Swiniarski, A. Skowron. "Rough set methods in feature selection and recognition" Pattern Recognition Letters.Vol. 24, pp. 833-849, 2003

[11] Pawlak, Z. and Skowron, A. "Rough sets: some extensions", Information Sciences, 177, 28-40, 2007.

[12] W. Ziarko. "Variable precision rough set model". Journal of Computer and System Science. Vol. 46, no. 1, pp. 39-59, 1993

[13] W. Zhu, "Generalized rough sets based on relations", Information Sciences, vol. 177 (22), pp. 4997–5011,2007.