International Journal of
Soft Computing and Software Engineering (JSCSE)

Vol.5, No.1, 2015
Published online: Jan 25, 2015

e-ISSN: 2251-7545
DOI: 10.7321/jscse.v5.n1.2

An Improved Feature Selection based on Neighborhood Positive Approximation Rough Set in Document Classification

[1] Mrs. Leena. H. Patil, [2] Dr. Mohammed Atique,
[*1,] *Research Scholar, Department of Computer Science and Engineering, Sant Gadge Baba Amravati University, Amravati, India*
[2,] *Associate Professor, Department of Computer Science and Engineering,*
Sant Gadge Baba Amravati University, Amravati, India
Email: [1]harshleena23@rediffmail.com, [2]mohd.atique@gmail.com

Abstract. Feature selection is a challenging problem in the field of machine learning, pattern recognition and data mining. Feature Subset Selection becomes an important preprocessing part in the area of data mining. In rough set theory, the problem of feature selection, called as attribute reduction, aims to retain the discriminatory power of original features. A large number of features is the problem in text categorization. Most of the features are noisy, redundant, relevant or irrelevant noise that can mislead the classifier and it may have different predictive power. Therefore, feature selection is often used in text categorization. It is most important to reduce dimensionality of the data to get smaller subset of features and relevant information within efficient computational time as time complexity is the major issue in feature selection. To deal with these problem many feature selection algorithms are available, still such algorithms are often computationally time consuming, and possess the problem of accuracy and stability. To overcome these problems we developed a framework based on neighborhood positive approximation rough set for feature subset selection in which the size of the neighborhood depends on the threshold value δ. In the proposed framework we obtain several representative and rank preservation of significance measures of attributes. In this paper firstly document preprocessing is performed. Secondly, a neighborhood positive approximation is used to accelerate the attribute reduction. Thirdly result validations based on classifiers are performed. Experimental results show that the improved feature selection based on neighborhood positive approximation rough set model becomes more efficient in terms of the stability, computational time and accuracy in dealing with large datasets.

Keywords: Introduction, document Preprocessing, Feature Selection, Rough set, Neighborhood positive approximation.

* ***Corresponding Author: Mrs. Leena H. Patil***
Research Scholar, Department of Computer Science and Engineering,
Sant Gadge Baba Amravati University, Amravati, India
Email: *harshleena23@rediffmail.com*

1. Introduction.

Now a day the number of text document on the internet is increasing tremendously. To deal with large amount of data, data mining becomes an important technology. Many data mining techniques are being used for extracting the valuable information such as categorizing, clustering, classification and analysis. Rapid growth of online information and document available in digitized form, text categorization becomes an important key technique for organizing the document. To organize the large amount of data and stored in a structured formats certain data mining techniques are able to use or extract the necessary information from the unstructured document collections. Because of this, text mining techniques are useful in processing these documents [1][2].

Feature Selection, called as attribute reduction is a common problem in the field of data mining, pattern recognition and machine learning. Recently both in number and dimensionality of items in dataset have grown drastically. In real world application databases, tens, hundreds and even thousands of attributes are stored. Features may be relevant or irrelevant it consist different predictive power. Mostly it is important to reduce dimensionality of the data to get smaller set of features and relevant information for decreasing the cost in storing and reduction the process time. To overcome the issues, few attributes to be omitted, this will not seriously affect on classification accuracy. Many techniques in feature selection have been categorized namely filter and wrapper. The former employs to selects attributes according to some significance measures such as consistency [4], information gain [3], distance [5], dependency [6] and others, later employs a learning algorithm to evaluate the attribute subsets. The significance measures are divided into major categories: distance based measure, consistency based measure, and entropy based measures. Rough set theory with attribute reduction offers a systematic framework for distance based measure which attempts to retain the ability of original features for the objects from the universe. Feature selection algorithm is divided into two categories: categorical attributes and numerical attribute. The former considers all features as categorical variables. While the latter all attributes are viewed as real value variables from the real world space. For numerical values two types of approaches are proposed; one relies on fuzzy rough set theory and other with discretization of numerical attribute [8]. To consistency with numerical attributes several approaches are developed. Features as granular instead of numerical attributes has introduced [19]. A dependency function in rough set framework to the fuzzy case and proposed QUICKREDUCT algorithm [10]. The concept of fuzzy rough set formed on computational domain which improves the computational efficiency introduced [11][12]. The attribute reduction algorithm takes all attributes to be a symbolic value in classical rough set theory. After preprocessing the originality, one can use classical rough set theory to select a subset of features in the last ten to twenty year many techniques of attribute reduction have been developed in rough set theory. Pawlaks' Rough set theory has proven to be the most effective tool for feature selection and knowledge discovery from categorical data in recent years[13][14][15][16]. A theoretic framework based on rough set theory called positive approximation which accelerates the algorithm of heuristic attribute reduction has proposed [17]. A general heuristic feature selection algorithm has developed [17][18]. The attribute reduction based on positive approximation is an effective accelerator and can efficiently obtain an attribute reduct [18]. The algorithms are still time consuming. A framework based on tolerance relation called positive approximation, used to accelerate heuristic algorithm for feature selection from incomplete data has introduced (IFSPA)[18]. Applying the idea of positive region reduction, a heuristic feature selection algorithm for incomplete decision table which keeps positive region of target decision unchanged [19]. The combinational entropy for measuring the uncertainty of an information system and used its conditional entropy to obtain feature subset [20]. To measure the uncertainty of an information system and apply the entropy to reduce redundant features [21]. The conditional entropy of Shannon's' is used to calculate the relative attribute reduction of a decision information system[22]. Several authors have used variants of Shannon's entropy or mutual information to measure uncertainty and construct heuristic algorithm of attribute reduction in rough set theory.

Document preprocessing is a process that extracts a set of new terms from the original document/ terms into some distinct key term set. Feature selection process a subset which selects from the original set based on some criteria of feature importance.

Each of the above idea preserves to a particular property of given decision table. However these methods are still time consuming and computationally very expensive, which are intolerable in dealing with large scale data sets of high dimensions. The main objective is to focus on how to preprocess the document and how to improve the computational time efficiently of an attribute reduction algorithm. In this paper a new Rough set framework based on neighborhood positive approximation have been developed. By integrating the concept of the representative and rank preservation, we construct a modified framework.

Considering neighborhood positive approximation an improved feature selection algorithm (NFSPA) for accelerating the computational time has been implemented. Experimental results shows the improved feature selection based on neighborhood positive approximation algorithm(NFSPA) retains the same attribute subset as that of the general neighborhood positive region algorithm (NPR). The improved algorithm maintains the stability with reduction in computational time. Two classifiers Naive Baise and KNN are used for result validation which shows the accuracy of feature subset selection. Rest of the paper is organized as: In Section 2 document preprocessing part has viewed briefly. Preliminaries of Neighborhood Rough set are briefly viewed in Section 3. In Section 4 neighborhood positive approximation and its properties are investigated. In Section 5, The Rank preservation of different representative significance measures on attributes are analyzed. In Section 6, we developed an improved feature selection algorithm based on neighborhood positive approximation algorithm (NFSPA). Section 7 includes the classifier Naive Baise and KNN based on classifier. An experimental analysis on three dataset Reuters 21578, Classic 04 and Newsgroup 20 is included in Section 8. Finally, Section 9 includes Conclusion.

2. Document Preprocessing

To categorize and browse thousands of documents smoothly, document preprocessing becomes a most important step. It describes the required transformation processing of documents to obtain the designated representation of documents. Thousands of words are present in a document set; the aim of this is to reduce dimensionality for having better accuracy for classification [23]. Document preprocessing is divided into following stages:

1. Stop words removing: Stop word List is used that contains the words to be excluded. The Stop word list is applied to remove terms that have a special meaning but do not discriminate for topics.
2. Word Stemming: Stemming algorithm such as porters is used to reduce a word to its stem or root form.
3. WordNet: WordNet Senses Disambiguation is applied as an English Database.
4. Global Unique words and frequent word set gets generated.

2.1 Stop-Word Removing

Stop-words are words that from non-linguistic view do not carry information. Stop-words remove the non-information behavior words from the text documents and reduce noisy data. Most existent search engines do not record stop-words in order to reduce the space and speed up the searches. To organize large corpus, removing the stop words affords the similar advantages. Firstly, it could save huge amount of space. Secondly, it helps to deduce the noises and keep the core words, and it will make later processing more effective and efficient.

2.2 Word Stemming

This process is used for transforming the words into their stem. In many languages the various syntactic form of words are used and explained. The most important technique called stemming is used for the reduction of words into their root. Many words from the English language can be reduced to their base form or stem e.g. agreed, agreeing, disagree, agreement and disagreement belong to agree. Porter Stemmer algorithm is a applied to stem documents. It is compact, simple and relatively accurate. It does not require creating a suffix list before applied.

2.3 Wordnet

WordNet is a linguistic English Database developed at cognitive science laboratory [26]of Princeton University. It has been developed with the aim of creating a machine tractable model of human language processing capability. It organizes words into a group called sysnets. Each of these contains a

e-ISSN: 2251-7545
DOI: 10.7321/jscse.v5.n1.2

collection of synonymous words and corresponds to a concept. Therefore, WordNet is considered to be an English thesaurus database which maps their concepts. WordNet is divided into four categories noun, verbs, adjectives and adverbs. Using lexical database the WordNet approach measures the relatedness of terms from the words. Based on these, we can compute scores of semantic relatedness of terms found from WordNet. In general as a dictionary, WordNet covers some specific terms from every subject related to their terms. Wordnet as a lexical database map all the stemmed words from the standard documents into their specifies lexical categories.

3. Preliminaries of Neighborhood Rough set

In this section, several basic concepts of rough set are reviewed. Throughout this paper we suppose that the universe data used is denoted by information system $(IS) = < U, A >$, where U is a non empty and finite set of samples $\{x_1, x_2, x_3, \dots, x_n\}$ called an universe. A is a set of attributes $\{a_1, a_{2,..,} a_n\}$ to characterize the samples. $< U, A >$ is called as decision table. If $A = C\ U\ D$, where C is a condition attribute and D is a decision attribute. For Example given an arbitrary variable $x_i \in U$ and $B \subseteq C$, the neighborhood $\delta_B(x_i)$ of x_i in feature space B is defined as

$\delta_B(x_i) = \{x_j \mid x_j \in \cup,\ \Delta^B(x_i, x_j) \leq \delta\}$,

where Δ is a distance function.

For $\forall_{x1,x2,x3} \in U$, it usually satisfies:

1. $\Delta(x_1, x_2) \geq 0$, $\Delta(x_1, x_2) = 0$ if and only if $x_1 = x_2$
2. $\Delta(x_1, x_2) = \Delta(x_1, x_2)$;
3. $\Delta(x_1, x_3) \leq \Delta(x_1, x_2) + \Delta(x_2, x_3)$;

The three different metric distance functions are most widely used in machine learning and pattern recognition.

A general metric names minkowsky distance can also be defined. Considered x_1 and x_2 as two objects in N- dimensional space $A = \{a_1, a_2, a_3, \dots, a_N\}$, $f(x, a_i)$ denotes the value of sample x in the ith attribute a_i. Therefore

$\Delta_P(x_1, x_2) = (\sum_{i=1}^{N} | f(x_i, a_i) - f(x_2, a_i)|^{P})^{1/P}$

Where (1) is called as Manhattan distance Δ_1 if $P = 1$; (2) Euclidean distance Δ_2, if $P = 2$. (3) Chebychev distance if $P = \infty$.

A detailed about distance function is explained in [30].

$\delta_B(x_i)$ is the neighborhood centered with sample x_i and the size of the neighborhood depends on the threshold δ. If δ has greater value then more samples gets connected into neighborhood of x_i. The shapes also depend on the used norm. A Neighborhood of x_0 in terms of different distances is as shown in Figure 1.

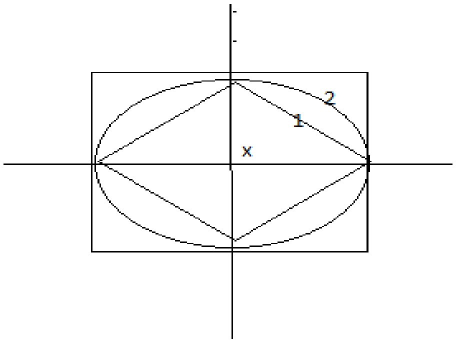

Figure 1. Neighborhoods for xo in terms of different distances.

Figure 1 shows the Rhombus region around center x_0 is manhattan distance. Ball Region is an Euclidean distance where as chebychev distance is a box region. Based on the discussion, two key factors play an important role to impact on the neighborhood i) used distance ii) threshold δ. The used distance determines the shape of neighborhood and threshold controls the size of the neighborhood. If $\delta = 0$ then neighborhood degrades to an equivalent class. Here the samples of the neighborhood granules are equivalent to each other whereas the samples of neighborhood rough set model degenerates to one. Hence, therefore the neighborhood rough set models are the natural generalization of the rough set. For example to deal to heterogeneous features, following definitions are used to compute Neighborhood samples with mixed attributes.

Let $B_1 \subseteq A\ and\ B_2 \subseteq A$ be numerical and categorical attribute. The neighborhood granules for sample x for $B_1, B_2\ and\ B_1 \cup B_2$are

(1) $\delta_{B_1}(x) = \{ x_i \mid \Delta_{B_1}(x, x_i) \le \delta, x_i \in U\}$;

(2) $\delta_{B_2}(x) = \{ x_i \mid \Delta_{B_2}(x, x_i) = 0, x_i \in U\}$;

(3) $\delta_{B_1 \cup B_2}(x) = \{ x_i \mid \Delta_{B_1}(x, x_i) \le \delta\,{}^\wedge\,\Delta_{B_2}(x, x_i) = 0, x_i \in U\}$; Where ∧ means "and" operator.

Where (1) is defined for numerical attributes, (2) is for categorical attributes and (3) is for mixed categorical and numerical attributes. Considering the above the samples for neighborhood granule have the same value for categorical data and distance for numerical feature is less than threshold $\boldsymbol{\delta}$. Besides these there are various distance functions for mixed numerical and categorical features[30][31]such as heterogeneous Euclidean overlap metric function (HEOM), value difference metric(VDM) and Heterogeneous value difference metric (HVDM). HEOM is defined as

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

where m is the number of attributes, w_{a_i} is the weight of attribute a_i, $d_{a_i}\left(x_{a_i}, y_{a_i}\right)$ is the distance between the samples x and y with respect to attribute a_i, is defined as

$$d_{a_i}(x, y) = \begin{cases} 1, \quad if\ the\ attribute\ value\ of\ x\ or\ y\ are\ unknown \\ overlap(x, y), if\ a_i\ is\ a\ categorical\ attribute \\ m\_diff_{a_i}(x, y), if\ a_i\ is\ a\ numerical\ attribute \end{cases}$$

Where $overlap(x, y) \begin{cases} 0, if\ x = y \\ 1\ otherwise \end{cases}\ and$

$$m_{diff_{a_i}}(x, y) = \frac{|\, x - y \,|}{max_{a_i} - min_{a_i}}$$

If the set of objects and the neighborhood relation N over U is called as neighborhood approximation space. For any $\boldsymbol{X} \subseteq \boldsymbol{U}$, two objects called lower and upper approximation of x in $< \boldsymbol{U}, \boldsymbol{N} >$ are defined as

$\underline{\boldsymbol{N}}\boldsymbol{X} = \{\boldsymbol{x_i} \mid \boldsymbol{\delta}(\boldsymbol{x_i}) \subseteq \boldsymbol{X}, \boldsymbol{x_i} \in \boldsymbol{U}\}$,

$\overline{\boldsymbol{N}}\boldsymbol{X} = \{\boldsymbol{x_i} \mid \boldsymbol{\delta}(\boldsymbol{x_i}) \cap \boldsymbol{X} \neq \emptyset, \boldsymbol{x_i} \in \boldsymbol{U}\}$

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

Example 1. A dataset consisting of numerical and categorical attribute given in Table 1. Where a is the numerical attribute, b the categorical attribute and D is the decision.

Vol.5, No.1, 2015
Published online: Jan 25, 2015

e-ISSN: 2251-7545
DOI: 10.7321/jscse.v5.n1.2

Initially compute the neighborhood samples with $\boldsymbol{\delta = 0.1}$. In this case an attribute a, $\boldsymbol{\delta(x_1) = \{x_1\}}$; , $\boldsymbol{\delta(x_2) = \{x_2, x_5\}}$; , $\boldsymbol{\delta(x_3) = \{x_3\}}$; , $\boldsymbol{\delta(x_4) = \{x_4, x_5\}}$; , $\boldsymbol{\delta(x_5) = \{x_2, x_4. x_5, x_6\}}$, $\boldsymbol{\delta(x_6) = \{x_4, x_5, x_6\}}$.
In this, samples can be divided into set of equivalence classes depending on the feature values of attribute b, $\boldsymbol{U/b = \{\{ x_1 x_2 x_5\} \{ x_3, x_4) \{ x_6\}}$. Considering the decision attribute, the samples are grouped into different subsets: $\boldsymbol{X_1 = \{ x_1, x_3, x_6)}$, $\boldsymbol{X_2 = \{ x_2, x_4, x_5\}}$. After approximation $\boldsymbol{X_1}$ with the granules induced by attribute a, we get $\boldsymbol{\underline{N}X_1 = \{x_1, x_2\}}$, $\boldsymbol{\overline{N}X_1 = \{ x_1, x_3, x_5, x_6 \}}$. Similarly $\boldsymbol{\underline{N}X_2 = \{x_2, x_4\}}$, $\boldsymbol{\overline{N}X_1 = \{ x_2, x_4, x_5, x_6 \}}$. Based on the preliminary, the information granules induced for a and b is listed as follow:

$$\boldsymbol{\delta(x_1) = \{x_1\} \cap \{ x_1, x_2, x_5\},}$$
$$\boldsymbol{\delta(x_2) = \{ x_2, x_5\},}$$
$$\boldsymbol{\delta(x_4) = \{ x_4),}$$
$$\boldsymbol{\delta(x_5) = \{x_2, x_5\},}$$
$$\boldsymbol{\delta(x_6) = \{ x_6\}}$$

With the information provided for attribute a and b, the lower and upper approximation of $\boldsymbol{X_1\ and\ X_2}$ are shown as follow:

$\boldsymbol{\underline{N}X_1 = \{ x_1, x_3, x_6\}}$, $\boldsymbol{\overline{N}\, X_1 = \{ x_1, x_3, x_6\}}$,
$\boldsymbol{\underline{N}X_2 = \{ x_2, x_4, x_5\}}$, $\boldsymbol{\overline{N}\, X_2 = \{ x_2, x_4, x_5\}}$

Table 1: Example of heterogeneous data

Object	A	b	D
1	0.20	1	N
2	0.85	1	Y
3	0.31	2	N
4	0.74	2	Y
5	0.82	1	Y
6	0.72	3	N

4. 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 $\mathrm{S}=< U, C \cup D, N >$ $\mathrm{X_1, X_2, \dots, X_N}$ are the object subsets with decisions 1 to N; δ_B (x_i) is the neighborhood information granule generated by attribute $\mathrm{B} \subseteq \mathrm{C}$, the lower and upper approximation of decision D with respect to attributes B are defined as

$\underline{\mathrm{N}}_\mathrm{B}\mathrm{D} = \cup_{\mathrm{i}=1}^{\mathrm{N}} \underline{\mathrm{N}_\mathrm{B}}\mathrm{X_i}$, $\overline{\mathrm{N}_\mathrm{B}}\mathrm{D} = \cup_{\mathrm{i}=1}^{\mathrm{N}} \overline{\mathrm{N}_\mathrm{B}}\mathrm{X_i}$,

Where

$\underline{\mathrm{N}}_\mathrm{B}\mathrm{X} = \{\mathrm{x_i} | \ \delta_\mathrm{B}\ (\mathrm{x}_\mathrm{i}) \subseteq \mathrm{X}, \mathrm{x_i} \in \mathrm{U}\}, \overline{\mathrm{N}_\mathrm{B}}\mathrm{X} = \{\mathrm{x_i} | \ \delta_\mathrm{B}\ (\mathrm{x}_\mathrm{i}) \cap \mathrm{X} \neq \emptyset, \mathrm{x_i} \in \mathrm{U}\}$

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

$\mathrm{BN(D)} = \overline{\mathrm{N}_\mathrm{B}}\mathrm{D} - \underline{\mathrm{N}_\mathrm{B}}\mathrm{D}$

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 $\mathrm{POS_B(D)}$. $\mathrm{POS_B(D)}$ is the subset of objects whose neighborhood granules consistently belong to one the decision classes. Figure 2 shows the positive, boundary and negative region.

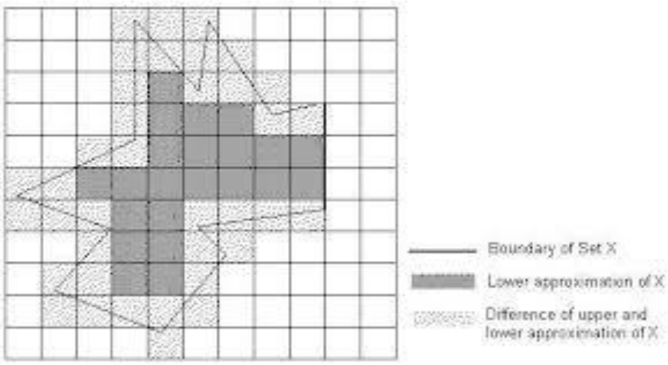

Figure 2.Rough set on positive, boundary and negative region

For different feature subspaces the classification task performs different boundary regions. The size of the boundary region reflects on the discernibly and recognition power of the corresponding subspace and condition attribute. As the boundary region changes or increases, the characterizing power of the condition attribute becomes weaker.

Definition 2: Given the decision table $S = < U, C \cup D >$, the distance function Δ and neighborhood size δ, the dependency degree of D to NP_i is defined as $\gamma_{NP_i}(D) = \frac{\left|POS^{U}_{NP_i}(D)\right|}{|U|}$
Where $|\cdot|$ is the cardinality of a set. $\gamma_{NP}(D)$ is the ability of NP to approximate D. As $POS_{NPi}\ (D) \subseteq U,$ 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. As it considers the categorical and numerical, one needs to proceed with preprocessing. To improve the searching strategy of forward greedy search algorithm, the concept of neighborhood positive region has been introduced. In this two important measures of attributes are used, namely inner importance measure and outer importance measure. The inner importance is used to determine the significance of each attribute whereas the outer importance measures are used for forward greedy search algorithm. 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 $S = < U, C \cup D >$, condition partition can be obtain as ${}^{U}/_{C} = \{X_1, X_2, \ldots, X_m\}$ and the decision partition as ${}^{U}/_{D} = \{Y_1, Y_2, \ldots, Y_n\}$. 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 $S = < U, C \cup D >, B \subseteq C\ and\ \forall a \in B$. 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 $S = < U, C \cup D >, B \subseteq C\ and\ \forall a\ \in C - B.$ 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 $S =< U, C\ \cup D >, B\ \subseteq A, \forall\ a\ \in B,$ 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 3.

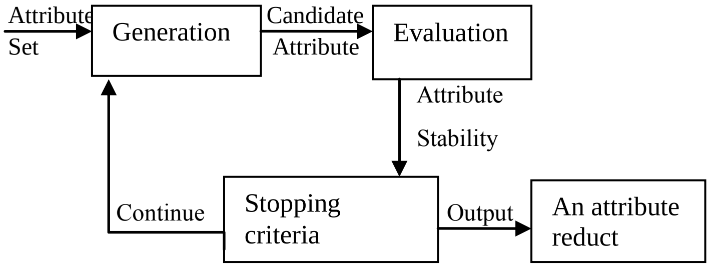

Figure 3: Forward greedy attribute reduction algorithm

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

e-ISSN: 2251-7545
DOI: 10.7321/jscse.v5.n1.2

This algorithm obtain the reduct attribute from the given decision table.

5. Feature Selection based on Neighborhood Positive Approximation (NFSPA)

A sequence of granulation worlds stretching from crude to fine granulation is determined as a sequence of attribute sets with granulation in the power set of attributes is called positive granulation world. In this a new set approximation approach called Neighborhood positive approximation has investigate the important properties by positive granulation world.

Given a decision table $\boldsymbol{S} =< \boldsymbol{U}, \boldsymbol{C}\ \boldsymbol{U}\ \boldsymbol{D} >, \frac{\boldsymbol{U}}{\boldsymbol{D}} = \{\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots\ \boldsymbol{Y_r}\}$ is called a target decision.

Definition 6: Let $\boldsymbol{S} = < \boldsymbol{U}, \boldsymbol{C} \cup \boldsymbol{D} >$ be a decision table, $\boldsymbol{X} \subseteq \boldsymbol{U}\ \boldsymbol{and}\ \boldsymbol{NP} = \{\boldsymbol{R_1}, \boldsymbol{R_2}, \ldots, \boldsymbol{R_n}\}$ a family of attribute sets with $\boldsymbol{R_1} \succcurlyeq \boldsymbol{R_2} \succcurlyeq \cdots \succcurlyeq \boldsymbol{R_n}\ (\boldsymbol{R_i} \in \boldsymbol{2^c})$. Given $\boldsymbol{NP_i} = \{\boldsymbol{R_1}, \boldsymbol{R_2}, \ldots, \boldsymbol{R_i}\}$, we define $\boldsymbol{NP_i}$- lower approximation sets $\underline{\boldsymbol{NP_i}}(\boldsymbol{X})$ and $\boldsymbol{NP_i}$- upper approximation sets $\overline{\boldsymbol{NP_i}}(\boldsymbol{X})$ of $\boldsymbol{NP_i}$ – neighborhood positive approximation of X as

$$\begin{cases} \underline{\boldsymbol{NP}}_{\boldsymbol{i}}(\boldsymbol{X}) = \bigcup_{\boldsymbol{k=1}}^{\boldsymbol{i}} \underline{\boldsymbol{R_k}}\ \boldsymbol{X_k} \\ \qquad \overline{\boldsymbol{NP}}_{\boldsymbol{i}}(\boldsymbol{X}) = \overline{\boldsymbol{R}}_{\boldsymbol{i}}\boldsymbol{X} \end{cases},$$

Where $\boldsymbol{X_1} = \boldsymbol{X}\ \boldsymbol{and}\ \boldsymbol{X_k} = \boldsymbol{X} - \bigcup_{\boldsymbol{j=1}}^{\boldsymbol{k=1}} \underline{\boldsymbol{R_j}}\, \boldsymbol{X_j}, \boldsymbol{k} = \boldsymbol{2}, \boldsymbol{3}, \ldots \boldsymbol{i},\ \boldsymbol{i} = \boldsymbol{1}, \boldsymbol{2}, \ldots \boldsymbol{n}.$

Correspondingly, the boundary of X is given as

$\boldsymbol{BN_{NP_i}}(\boldsymbol{X}) = \overline{\boldsymbol{NP_i}}(\boldsymbol{X}) - \underline{\boldsymbol{NP_i}}(\boldsymbol{X}).$

Theorem 1. Let $\boldsymbol{S} =< \boldsymbol{U}, \boldsymbol{C} \cup \boldsymbol{D} >$ be a decision table, $\boldsymbol{X} \subseteq \boldsymbol{U}\ \boldsymbol{and}\ \boldsymbol{NP} = \{\boldsymbol{R_1}, \boldsymbol{R_2}, \ldots, \boldsymbol{R_n}\}$ a family of attribute sets with $\boldsymbol{R_1} \succcurlyeq \boldsymbol{R_2} \succcurlyeq \cdots \succcurlyeq \boldsymbol{R_n}\ (\boldsymbol{R_i} \in \boldsymbol{2^C})$.

Given $\boldsymbol{NP_i} = \{\boldsymbol{R_1}, \boldsymbol{R_2}, \ldots, \boldsymbol{R_i}\}, \boldsymbol{then}\ \forall \boldsymbol{NP_i}(\boldsymbol{i} = \boldsymbol{1}, \boldsymbol{2}, .., \boldsymbol{n})$, we have $\underline{\boldsymbol{NP_i}}(\boldsymbol{X}) \subseteq \boldsymbol{X} \subseteq \overline{\boldsymbol{NP_i}}(\boldsymbol{X})$,
$\underline{\boldsymbol{NP_1}}(\boldsymbol{X}) \subseteq \underline{\boldsymbol{NP_2}}(\boldsymbol{X}) \subseteq \cdots \subseteq \underline{\boldsymbol{NP_i}}(\boldsymbol{X})$

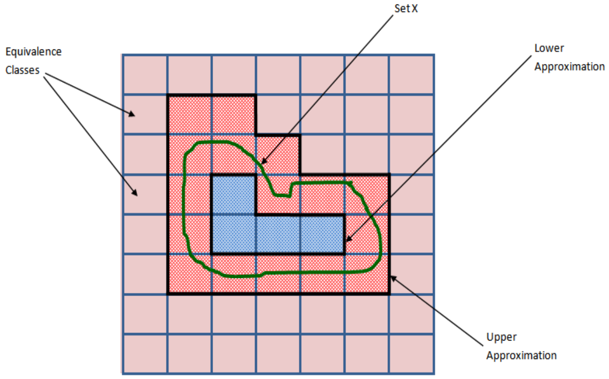

Figure 4. Visualizes the mechanism of neighborhood positive approximation.

Definition 7: Let $\boldsymbol{S} = < \boldsymbol{U}, \boldsymbol{C} \cup \boldsymbol{D} >$ be a decision table, $\boldsymbol{NP_i} = \{\boldsymbol{R_1}, \boldsymbol{R_2}, \ldots, \boldsymbol{R_n}\}$ a family of attribute sets with $\boldsymbol{R_1} \succcurlyeq \boldsymbol{R_2} \succcurlyeq \cdots \succcurlyeq \boldsymbol{R_n}\ (\boldsymbol{R_i} \in \boldsymbol{2^c})$ and $\frac{\boldsymbol{U}}{\boldsymbol{D}} = \{\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots\ \boldsymbol{Y_r}\}$.

Lower approximation and upper approximation of D with respect to $\boldsymbol{NP_i}$ are defined as

$$\begin{cases} \underline{\boldsymbol{NP_i}}(\boldsymbol{D}) = \{\underline{\boldsymbol{NP_i}}(\boldsymbol{Y_1}), \underline{\boldsymbol{NP_i}}(\boldsymbol{Y_2}), \ldots., \underline{\boldsymbol{NP_i}}(\boldsymbol{Y_r})\} \\ \boldsymbol{N}\overline{\boldsymbol{P_i}}(\boldsymbol{D}) = \{\overline{\boldsymbol{NP_i}}(\boldsymbol{Y_1}), \overline{\boldsymbol{NP_i}}(\boldsymbol{Y_2}), \ldots., \overline{\boldsymbol{NP_i}}(\boldsymbol{Y_r})\} \end{cases}$$

$\underline{\boldsymbol{NP_i}}\boldsymbol{D}$ is called the neighborhood positive region of D with respect to the granulation order $\boldsymbol{NP_i}$, denoted by $\boldsymbol{POS_{NP_i}^U}(\boldsymbol{D}) = \bigcup_{\boldsymbol{k=1}}^{\boldsymbol{r}} \boldsymbol{NP_i Y_k}.$

Theorem 2: Let $S = < U, C \cup D >$ be a decision table, $X \subseteq U\ and\ NP = \{ R_1, R_2, \dots, R_n\}$ a family of attribute sets with $R_1 \succcurlyeq R_2 \succcurlyeq \cdots \succcurlyeq R_n$ $(R_i \in 2^c)$. Given $NP_i = \{ R_1, R_2, \dots, R_i\}$, we have
$POS^U_{NP_{i+1}}(D) = POS^U_{NP_i}(D) \cup POS^{U_{i+1}}_{NP_{i+1}}(D)$
Where $U_1 = U\ and\ U_{i+1} = U - POS^U_{NP_i}(D)$.
Example 2: Let $S =< U, C \cup D >$ be a decision table, where $U = \{h_1, h_2, h_3, h_4, h_5, h_6, h_7, h_8\}, C = \{ a_1, a_2\}$,
$\frac{U}{D} = \{\{h_1, h_2, h_3, h_4, h_7, h_8\}, \{h_5, h_6\}\}\ and$
$\frac{U}{\{a_1\}} = \{\{h_1\}, \{h_2\}, \{h_3, h_4\}, \{h_5, h_6, h_7, h_8\}\}$,
$\frac{U}{C} = \{\{h_1\}, \{h_2\}, \{h_3, h_4\}, \{h_5, h_6\}, \{h_7, h_8\}\}$
be two partitions on U. Obviously, $\{a_1\} \succcurlyeq C$ holds.. Thus, one can construct two granulation orders, $NP_1 = \{a_1\}\ and\ NP_2 = \{a_1, a_2\}$.
By computing the neighborhood positive approximation of D, one can easily obtain an $POS^U_{NP_1}(D) = POS^{U_1}_{NP_1}(D) = \{h_1, h_2, h_3, h_4\}$,
Where $U_1 = U, U_2 = U - POS^U_{NP_1}(D) = \{ h_5, h_6, h_7, h_8\}$,
$POS^{U_2}_{NP_2}(D) = \{h_5, h_6\}$
Hence
$POS^U_{NP_2}(D) = \{h_1, h_2, h_3, h_4, h_5, h_6\} = POS^U_{NP_1}(D) \cup POS^{U_2}_{NP_2}(D)$

This mechanism implies for improving the computing performance of attribute reduction algorithm.

Definition 8: Let the decision table $S =< U, C \cup D >, B \subseteq A$, we say attribute set B is a relative reduct if
1) $\gamma_B(D) = \gamma_A(D)$;
2) $\forall a \in B, \gamma_B(D) > \gamma_{B-a}(D)$.
The first condition says that $POS_B(D) = POS_A(D)$ and the second condition says that there is no superfluous attribute in the reduct. Hence a reduct contains the minimal subset of attribute which has the same power as the whole attribute set.
To improve the searching strategy of an attribute reduction algorithm, we introduce the rank preservation of the significance measures of the attribute based on the neighborhood positive approximation from the decision table. The rank preservation of the significance attribute is based on the dependency measure and is denoted by $SIG^{outer}_{\Delta} = (a, B, D, U)$
From the above this can be proved as:

Definition 9: Let decision table $S =< U, C \cup D >, B \subseteq C\ and\ U^1 = U - POS^U_B(D).\ For\ \forall a, b \in C - B,\ if\ SIG^{outer}_1(a, B, D, U) \geq SIG^{outer}_1(b, B, D, U),\ then\ SIG^{outer}_1(a, B, D, U^1) \geq SIG^{outer}_1(b, B, D, U^1)$.

Proof: From the definition of $SIG^{outer}_1(a, B, D) = \gamma_{B\cup a}(D) - \gamma_B(D)$. In this the value only depends on the dependency function $\gamma_B(D) = \frac{|POS_B(D)|}{|U|}$, since $U^1 = U - POS^U_B(D)$.
we know that $POS^U_B(D) = \emptyset\ and\ POS^{U^1}_{B\cup a}(D) = POS^{U^1}_{B\cup a} = POS^U_{B\cup a}(D) - POS^U_B(D)$.
Therefore we have,

$$\frac{SIG^{outer}_1(a,B,D,U)}{SIG^{outer}_1(a,B,D,U^1)} = \frac{\gamma^U_{B\cup a}(D) - \gamma^U_B(D)}{\gamma^{U^1}_{B\cup a}(D) - \gamma^{U^1}_B(D)} = \frac{|U^1|\,|POS^U_{B\cup a}(D) - POS^U_B(D)|}{|U|\,\left|POS^{U^1}_{B\cup a}(D) - POS^{U^1}_B(D)\right|} = \frac{|U^1|\,|POS^U_{B\cup a}(D) - POS^U_B(D)|}{|U|\,|POS^U_{B\cup a}(D) - POS^U_B(D)|}$$
$$= \frac{|U^1|}{|U|}$$

Because $\frac{|U^1|}{|U|} \geq \mathbf{0}\ and\ if\ SIG_1^{outer}(a,B,D,U) \geq SIG_1^{outer}(b,B,D,U),\ then\ SIG_1^{outer}(a,B,D,U^1) \geq SIG_1^{outer}(b,B,D,U^1)$. Hence it completes the proof.

Algorithm 2: An improved Feature Selection Neighborhood positive approximation algorithm(NFSPA)

```
Input     1. Decision Table S=<U, C ∪ D >
          2. Delta $\delta$ // $\delta$ is the threshold to control the size of neighborhood.
Output:   One reduct red
Step 1:   $red \leftarrow \emptyset$ ; //red is the pool to conserve the selected attribute.
Step 2:   For each $a_k \in C - red$
Step 3:   Compute $\gamma_{red \cup a_i}(D) = \frac{|POS_{B \cup a_i}(D)|}{|U|}$
Step 4:   Compute $Sig_1^{inner}(a,B,D) = \gamma_B(D) - \gamma_{B-\{a\}}(D)$
Step 5:   Compute $Sig_1^{outer}(a,B,D) = \gamma_{B \cup \{a\}}(D) - \gamma_B(D)$
Step 6    end
Step 7:   Compute $Sig^{inner}(a_k, C, D, U), k \leq |C|$
Step 8:   Put $a_k$ into red where $Sig^{inner}(a_k, C, D, U) > 0$; // these attributes are the core of the given decision table.
Step 9:   $i = 1$; $R_1 = red$; $P_1 = \{R_1\} and\ U_i \leftarrow U$
Step 10:  While $EF^{U_i}(red, D) \neq EF^{U_i}(C, D)\ do$
{
    Compute positive region of neighborhood positive approximation
    $POS_{NP_i}^{U}(D)$,
    $U_i = U - POS_{NP_i}^{U}(D)$
    $i \leftarrow i + 1$; $red \leftarrow red \cup \{a_0\}$, where
            $Sig^{outer}(a_0, red, D) = \max\{Sig^{outer}(a_k, red, D),\quad a_k \in C - red$
    $R_i \leftarrow R_i \cup \{a_0\}$,
    $NP_i \leftarrow \{R_1, R_2, \dots, R_i\}$
};
Step 11:  Return red
Step 12:  end.
```

Computing the significance measure of an attribute from step 6, the time complexity becomes$O(|U|)$. Therefore the time complexity for core attributes in step 6 is $O(|C||U|)$. In step 9 we consider the core attributes and keep on adding the significance attribute into each set till finding the reduct. It is so called as forward attribute reduction algorithm where the time complexity is $O(\sum_{i=1}^{|C|}|U_i|(|C| - i + 1))$ and therefore the time complexity of feature selection based on neighborhood positive approximation (FSNPA) becomes $O(|U||C| + \sum_{i=1}^{|C|}|U_i|(|C| - i + 1))$. However the time complexity of general attribute reduction algorithm is $O(|U||C| + \sum_{i=1}^{|C|}|U|(|C| - i + 1))$. Hence the time complexity of FSNPA becomes much lower as compared to the general attribute reduction algorithm (NPR)and also reduces the computation time.

6. Result Validation based on classifiers

6.1 Classifier

As the size of information grows rapidly, the problem arises of handling the data. It is infeasible to classify the data manually so automatic methods have been approached to reduce the time and effort for classification. Many document classifications has been built to categories the document according to their content. To improve the accuracy of classifier, researchers have worked on many ranking methods which select the term such as term frequency, chi squared, mutual information, and information gain. Still the problem arises is redundancy in the selected term. Redundant terms are equivalent to noise which causes a reduction in the accuracy of classifier. The classification accuracy changes according to the features being input to the classifier. If the features are of less redundant then the accuracy increases else it decreases. Feature selection algorithm with redundancy reduction for text classification, algorithm helps to decrease in redundant which improves the efficiency of the classifier[89].

6.2 Naive Bayes

Most Widely used classifier is the naive bayes. This classifier built the concept of probabilistic Classification where the probability is calculated for each document. it shows the belonging to the categories specified [10][12][13]. Many approaches using naive bayes classifier, multinomial naive bayes is used where the probability $P(C_j \mid d_i)$ of a document d_i belongs to the category C_j. C_j is calculated through the following equation :

$$P(C_j|d_i) = \frac{P(C_j) \prod_{k=1}^{|d|} P(w_{di,k}|C_j)}{\sum_{r=1}^{|c|} P(C_r) \prod_{k=1}^{|dj|} P(w_{d,k}|C_r|)}$$

where $|C|$ is the number of categories. $|d_i|$ be the length of document. $P(C_j)$ is probability of category is calculated according to the equation.

$$P(C_j) = \frac{1 + \sum_{i=1}^{|D|} P(C_j|d_i)}{|C| + |D|}$$

The probability of word gives that the category occurred $P\ (w_i \mid c_j)$ is calculated through the equation.c

$$P(w_i|c_j) = \frac{1 + \sum_{i=1}^{|D|} N(w_i, d_i)\, P\ (\ y_i = c_j|d_i)}{|v| + \sum_{i=1}^{|D|} \sum_{j=1}^{|D|} N(\ w_{i,} d_i)\, P(y_i = c_j|d_i)}$$

where $|D|$ is the number of documents in the training set, $|v|$ is the number of words in the training set.

6.3 K-Nearest Neighbor

The KNN [5] algorithm is a well-known instance-based approach that has been widely applied to text categorization due to its simplicity and accuracy [7][9]. To categorize an unknown document, the KNN classifier ranks the document's neighbors among the training documents and uses the class labels of the k most similar neighbors. Similarity between two documents may be measured by the Euclidean distance, cosine measure, etc. The similarity score of each nearest neighbor document to the test document is used as the weight of the classes of the neighbor document. If a specific category is shared by more than one of the k-nearest neighbors, then the sum of the similarity scores of those neighbors is obtained from the weight of that particular shared category [2]. A detailed procedure of KNN can be referred to in Cover and Hart [5].

When classification is done by means of the KNN, the most important parameter affecting classification is k-nearest neighbor number. Usually, the optimal value of k is empirically determined. k value is determined so that it would give the least classification error.

e-ISSN: 2251-7545
DOI: 10.7321/jscse.v5.n1.2

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. We also 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 2. 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 2: Data set description

Sr.No	Data Set	Cases	Features	Classes
01	Reuters 21578	212	6539	04
02	Classic 4	54	1625	06
03	Newsgroup 20	52	1454	04

The document preprocessing is performed in four stages. the first step is of removing the stop words. These stopwords are removed because those are useless for classification. In this the stop words are removed according the existing stop word list 571 words. After removal of stopwords, the dataset consist of global unique words. In the second stage, the porter stemming algorithm is applied.. In the third stage we use wordnet to find the terms approach measures the relatedness of terms from the words. In the fourth stage rare word technique is applied. In this we prune the words that appear less two times in the documents. After applying the pruning technique, the total number of features is finally extracted. The details are shown in Table 3.

Table 3: Preprocessed Document

Sr.No	Data Set	No. of Documents	Features extracted
01	Reuters 21578	212	5677
02	Classic 4	54	1411
03	Newsgroup 20	52	976

Experimental results shows that the number of features selected for NPR and Improved NFSPA are same. Hence the stability of Improved NFSPA algorithm retains same the predictive power of original features. The computational time for Improved NFSPA largely reduces as compare to the general NPR algorithm. The details of computational time and number of feature selection for different threshold values are shown in Table 4-6. 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 3-5 shows the computational time with the two algorithms NPR and NFSPA. Each one present computational time of NPR and NFSPA for delta values 0.01, 0.015 and 0.001. In these the computational time with NFSPA significantly reduces with same number of features.

Table 5: Time and feature selection of the algorithms NPR and improved NFSPA with $\delta = 0.01$

Sr.No	Data Set	Features	NPR		NFSPA	
			Features	Time (s)	Features	Time (s)
01	Reuters 21578	5677	465	74.00	465	60.00
02	Classic 4	1411	101	40.00	101	25.00
03	Newsgroup 20	976	76	25.00	76	15.00

Table 6: Time and feature selection of the algorithms NPR and improved NFSPA with$\delta = 0.015$

Sr.No	Data Set	Features	NPR		NFSPA	
			Features	Time (s)	Features	Time (s)
01	Reuters 21578	5677	448	60.00	448	55.00
02	Classic 4	1411	92	30.00	92	22.00
03	Newsgroup 20	976	71	20.00	71	10.00

Table 7: Time and feature selection of the algorithms NPR and improved NFSPA with$\delta = 0.001$

Sr.No	Data Set	Features	NPR		NFSPA	
			Features	Time (s)	Features	Time (s)
01	Reuters 21578	5677	441	50.00	441	45.00
02	Classic 4	1411	86	25.00	86	20.00
03	Newsgroup 20	976	65	15.00	65	08.00

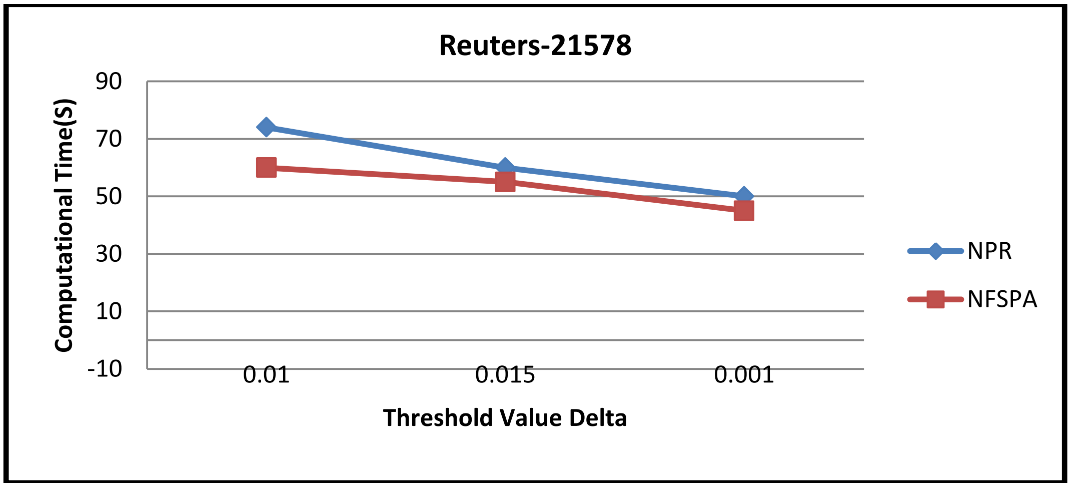

Figure 3: Comparison of computational time on Reuters 21578

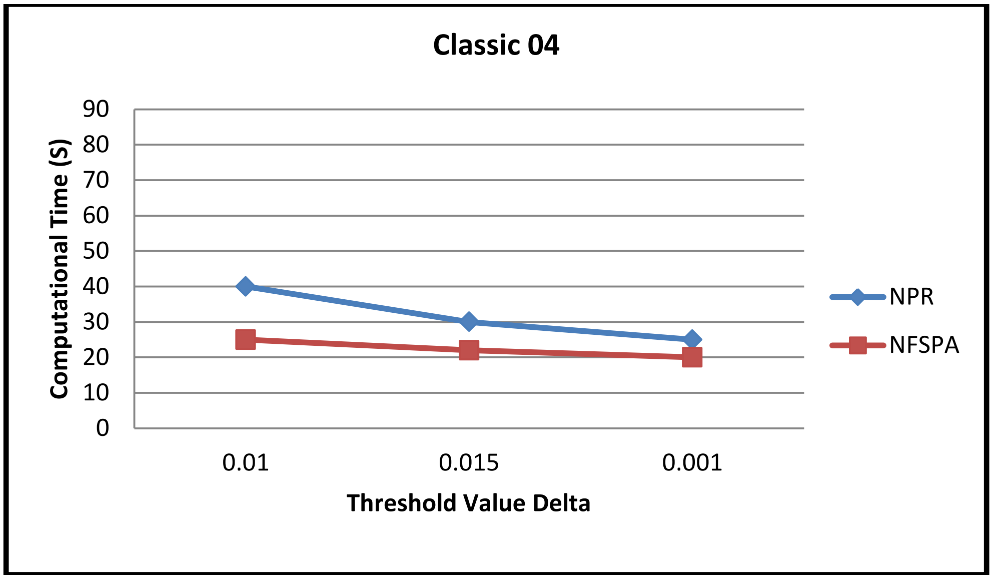

Figure 4: Comparison of computational time on Classic 04

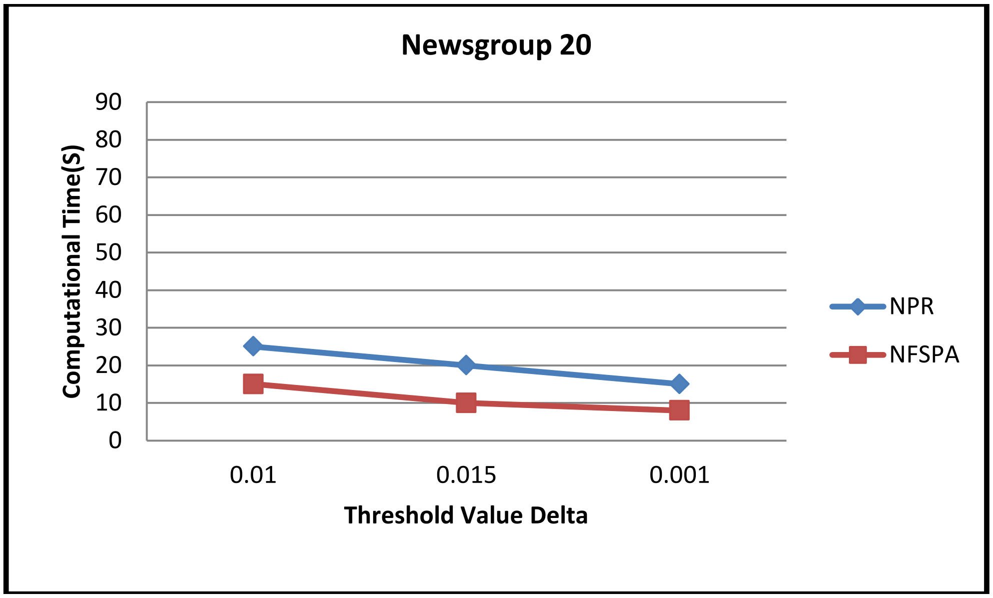

Figure 5: Comparison of computational time on Newsgroup 20

For result validation of general NPR algorithm and improved NFSPA algorithm, we used Naïve Baise and KNN classifiers. This classifier shows the accuracy of feature selection algorithm. The accuracy for general NPR algorithm and improved NFSPA algorithm retains same which shows that there is no loss of relevant information even after modification in general algorithm. The accuracy results for three data sets and feature selection algorithms with Naïve Baise and KNN classifier is shown in Table 7 and Table 8.

Table 7: The percentage accuracy of selected features with Naïve Baise and KNN classifier.

Sr. No	Data set	Features	Percentage Accuracy	
			NB	KNN
01	Reuters 21578	5677	97.85	98.85
02	Classic 4	1411	96.14	97.36
03	Newsgroup 20	976	92.45	94.75

Table 8: The percentage accuracy of selected features with Naïve Baise and KNN classifier.

Sr. No	Data Set	Features	NPR		Improved Algorithm(FSNPA)	
			NB	KNN	NB	KNN
01	Reuters 21578	465	97.85	98.85	97.85	98.85
02	Classic 4	101	96.14	97.36	96.14	97.36
03	Newsgroup 20	76	92.45	94.75	92.45	94.75

8. Conclusion

To overcome the limitation of existing feature selection algorithm, a theoretic framework based distance relation have been proposed in this study, which is called as neighborhood positive approximation rough set. This can be used to reduce computational time for feature selection in dealing with large data sets. Based on this framework, improved neighborhood feature selection positive approximation rough set (NFSPA) has been presented. For feature selection several representative algorithms have 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 neighborhood positive approximation rough set model is more efficient concern with the stability, computational time and accuracy in dealing with large datasets.

References

[1] Chun-Ling Chen a, Frank S.C. Tseng b, Tyne Liang, An integration of WordNet and fuzzy association rule mining for multi-label document clustering, Data & Knowledge Engineering 69 ,1208–1226, 2010.

[2] F. Beil, M. Ester, X. Xu, Frequent term-based text clustering, Proc. of Int'l Conf. on knowledge Discovery and Data Mining KDD '02, pp. 436–442, 2002

[3] C.K. Lee, G.G. Lee, Information gain and divergence-based feature selection for machine learning-based text categorization, Information Processing Manage 42 155–165, 2006

[4] M. Dash, H. Liu, Consistency-based search in feature selection, Artificial Intelligence 151 155–176, 2003.

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

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

e-ISSN: 2251-7545
DOI: 10.7321/jscse.v5.n1.2

[7] 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.

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

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

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

[11] R.B. Bhatt, M. Gopal, On fuzzy-rough sets approach to feature selection, Pattern Recognition Letters 26, 965–975, 2005

[12] R.B. Bhatt, M. Gopal, On the compact computational domain of fuzzy-rough sets, Pattern Recognition Letters 26 1632–1640, 2005.

[13] D.G. Chen, C.Z. Wang, Q.H. Hu, A new approach to attribute reduction of consistent and inconsistent covering decision systems with covering rough sets, Information Sciences 177, 3500 – 3518, 2007

[14] Q.H. Hu, X.D. Li, D.R. Yu, Analysis on classification performance of rough set based reducts, in: Q. Yang, G. Webb (Eds.), PRICAI 2006, LNAI, vol. 4099, Springer-Verlag, Berlin, Heidelberg, pp. 423 – 433, 2006.

[15] Z. Pawlak, Rough Sets, Theoretical Aspects of Reasoning About Data, Kluwer Academic Publishers, Dordrecht, 1991.

[16] Z. Pawlak, A. Skowron, Rough Sets: Some Extensions, Information Sciences 177, 28 – 40, 2007

[17] 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

[18] 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

[19] 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

[20] Y.H. Qian, J.Y. Liang, F. Wang, A new method for measuring the uncertainty in incomplete information systems, Fuzziness and Knowledge-Based Systems 17 (6), 855–880, 2009.

[21] J.Y. Liang, K.S. Chin, C.Y. Dang, C.M. Yam Richid, A new method for measuring uncertainty and fuzziness in rough set theory, International Journal of General Systems 31 (4) 331–342, 2002

[22] G.Y. Wang, H. Yu, D.C. Yang, Decision table reduction based on conditional information entropy, Chinese Journal of Computer 25 (7) 759–766, 2002.

[23] C.L.Chen, F.S.C. Tseng, T. Liang, An integration of fuzzy association rules and WordNet for document clustering, Proc. of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 147–159, 2009.

e-ISSN: 2251-7545
DOI: 10.7321/jscse.v5.n1.2

[24] Porter, M. F. An algorithm for suffix stripping. Program, 14(3), 130 – 137, 1980.

[25] George A. Miller –"WordNet: A Lexical Database for English".

[26] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Boston, 1991.

[27] Z. Pawlak, A. Skowron, Rudiments of rough sets, nformation Sciences 177 (1) 3–27, 2007.

[28] Z. Pawlak, A. Skowron, Rough sets: some extensions, Information Sciences 177, 28–40, 2007.

[29] D. Randall Wilson, Tony R. Martinez, Improved heterogeneous distance functions, Journal of Artificial Intelligence Research 6, 1 – 34, 1997.

[30] H. Wang, Nearest neighbors by neighborhood counting, IEEE Transactions on PAMI 28 942 – 953, 2006