An Exponential Kernel based Fuzzy Rough Sets Model for Feature Selection

Riaj Uddin Mazumder
Department of Mathematics
Assam University
Silchar

Shahin Ara Begum
Department of Computer Science
Assam University
Silchar

Devajyoti Biswas
Department of Mathematics
Assam University
Silchar

ABSTRACT
Feature subset selection is a data preprocessing step for pattern recognition, machine learning and data mining. In real world applications an excess amount of features present in the training data may result in significantly slowing down of the learning process and may increase the risk of the learning classifier to over fit redundant features. Fuzzy rough set plays a prominent role in dealing with imprecision and uncertainty. Some problem domains have motivated the hybridization of fuzzy rough sets with kernel methods. In this paper, the Exponential kernel is integrated with the fuzzy rough sets approach and an Exponential kernel approximation based fuzzy rough set method is presented for feature subset selection. Algorithms for feature ranking and reduction based on fuzzy dependency and exponential kernel functions are presented. The performance of the Exponential kernel approximation based fuzzy rough set is compared with the Gaussian kernel approximation and the neighborhood rough sets for feature subset selection. Experimental results demonstrate the effectiveness of the Exponential kernel based fuzzy rough sets approach for feature selection in improving the classification accuracy in comparison to Gaussian kernel approximation and neighborhood rough sets approach.

Keywords
Rough set, Fuzzy rough set, Exponential kernel, Feature selection

1. INTRODUCTION
Kernel methods in machine learning allow mapping the data into a high dimensional feature space in order to increase the classification power of linear learning algorithms [1], [2], [3]. Rough set theory proposed by Pawlak [4] has proven to be an effective tool for feature selection, knowledge discovery and rule extraction from categorical data. Rough set theory can also be used to deal with both vagueness and uncertainty in data sets. The rough set theory has some difficulty in handling both symbolic and real-valued values of attributes [5]. Rough set theory and fuzzy set theory are combined together [6], [7], [8] to deal with numeric and fuzzy attributes in an information system and both are complementary. Dubois and Prade [6] first proposed the concept of fuzzy rough sets. Fuzzy rough sets offer a high degree of flexibility in enabling the vagueness and imprecision present in real-valued data to be modeled effectively.

Most of the fuzzy rough sets are established based on fuzzy granules induced by fuzzy T-equivalence relation. It has been shown that kernel matrix computed with a reflexive kernel taking values from the unit interval [0, 1] is a fuzzy T-equivalence relation [9], [10]. Therefore, it is desiarable to consider such kernel functions to induce fuzzy T-equivalence relations from data. Exponential kernel functions are reflexive and symmetric in the unit interval [0, 1].

In this paper an Exponential kernel based fuzzy rough set for feature selection based on the properties discussed in [6], [7], [8], [11] is presented. The kernel functions extract fuzzy relations from data into fuzzy rough sets.

Rest of the paper is organized as follows: Section 2 presents a brief introduction to fuzzy rough sets and kernel methods. In section 3, an Exponential kernel based fuzzy rough set model for feature selection is introduced. In Section 4, feature selection with Exponential kernel based fuzzy rough sets is presented. Section 5 presents the experimental results. The paper is concluded in Section 6.

2. BACKGROUND
This section presents a brief introduction to fuzzy rough sets and kernel methods.

2.1 Fuzzy-Rough Sets
Fuzzy set theory and Rough set theory complement each other and as such constitute important components of soft computing. Researchers have explored a variety of different ways in which these two theories can interact with each other. There are many possibilities for rough-fuzzy hybridization; the most typical ones are to fuzzify sets to be approximated and/or to fuzzify the equivalence relation in an approximation space [12]. The first case allows obtaining rough approximations of fuzzy sets which results in the rough- fuzzy sets; while the second case allows obtaining approximations of fuzzy sets by means of fuzzy similarity relations resulting in the fuzzy- rough sets.

In the context of rough set theory [5], an equivalence relation is a fundamental and primitive notion. For fuzzy-rough sets, a fuzzy similarity relation is used to replace an equivalence relation. Let U be a nonempty universe, for a given fuzzy set A and a fuzzy partition $\Phi = \{F_1, F_2, \ldots, F_n\}$ on the universe U, the membership functions of the lower and upper approximations of A by Φ are defined as follows:

$$\mu_{\underline{\Phi}(A)}(F_i) = inf_x I\left(\mu_{F_i}(x), \mu_A(x)\right)$$

and

$$\mu_{\overline{\Phi}(A)}(F_i) = sup_x T\left(\mu_{F_i}(x), \mu_A(x)\right)$$

where, T and I denote a T-norm operator and implicator. The pair of sets $\langle \underline{\Phi}(F), \overline{\Phi}(F) \rangle$ is called a fuzzy-rough set. A general study of fuzzy-rough sets from the constructive and the axiomatic approaches is presented by Yeung *et al*. [8]. There is significant theoretical work on hybridization of fuzzy and rough set theories, as well as its usage in classification and similar supervised learning techniques [13].

2.2 Kernel methods

The challenges of machine learning have received much attention by the use of kernel methods. Kernel methods allows mapping the data into a high dimensional feature space in order to increase the computation of linear learning algorithms [3]. Kernel defines a similarity measure between two data points and allows the utilization of prior knowledge of the problem domain. Kernel provides all of the information about the relative positions of the inputs in the feature space so that the associated learning algorithm is based only on the kernel function. In the statistical perspective, Symmetric positive definite functions are called covariances. Hence kernels are essentially covariance based. In general there are two important classes of kernels, *viz*. stationary and non-stationary kernels [14].

(i) Stationary kernels: A stationary kernel is one which is translation invariant:

$$K(x, y) = K_S(x - y),$$

it depends only on the lag vector separating the two objects x and y, but not on the objects themselves. To emphasize the dependence on both the direction and the length of the lag vector, it sometimes called as an isotropic stationary kernel. Thus, a stationary kernel depends only on the norm of the lag vector between two objects and not on the direction, then the kernel is said to be isotropic (or homogeneous), and is thus only a function of distance and its covariance form is

$$K_{cov}(x, y) = K_I(\|x - y\|),$$

And the correlation form representation is:

$$K_{cor}(x, y) = K_I(\|x - y\|) \,/\, K_I(0),$$

Some commonly used isotropic stationary kernels are given in the table 1.

Table 1: Isotropic stationary kernels [14]

Name of Kernel	$K_I(\|x - y\|) \,/\, K_I(0)$
1) Rational quadratic positive definite in $\mathbb{R}^n$	$K(x, y) = 1 - \frac{\|x - y\|^2}{\|x - y\|^2 + \theta}$
2) Exponential positive definite in $\mathbb{R}^n$	$K(x, y) = \exp\left(-\frac{\|x - y\|}{\theta}\right)$
3) Gaussian positive definite in $\mathbb{R}^n$	$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{\theta}\right)$
4) Wave positive definite in $\mathbb{R}^3$	$K(x, y) = \frac{\theta}{\|x - y\|} \sin\left(\frac{\|x - y\|}{\theta}\right)$
5) Spherical positive definite in $\mathbb{R}^3$	$1 - \frac{3}{2}\frac{\|x-y\|}{\theta} + \frac{1}{2}\left(\frac{\|x-y\|}{\theta}\right)^3$ if $\|x - y\| < \theta$; zero otherwise

Further for each class of kernels, one can view their spectral representation and show how it can be used to design many new kernels.

(ii) Non-stationary kernels: The most general class of kernels is the one of non-stationary kernels depend explicitly on the two objects x and y such that $K(x, y) = (x^T y)^q$, q is the polynomial kernel degree. The reflexivity property holds on specific non-stationary kernels. For example in $\mathbb{R}^2$, the nonstationary kernel defined by [14]:

$$K(x, y) = \frac{\|x\| + \|y\| - \|x - y\|}{2\sqrt{\|x\|\|y\|}}$$

is reflexive and is stationary reducible.

3. EXPONENTIAL KERNEL BASED FUZZY ROUGH SET MODEL

In this section the Exponential kernel for computing fuzzy T-equivalence relations in fuzzy rough sets is introduced.

Let U be a non empty finite set (universe of discourse) samples, x_i is contained in U and is described by a vector $x_{ij} \in R^n$, where $j = 1, 2, \ldots, n$. Thus, $U \subseteq R^n$.
The Exponential kernel [14] is defined as:

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|}{\theta}\right),$$

where, $\|x_i - x_j\|$ is the Euclidiean distance between samples x_i and x_j; and (i) $K(x_i, x_j) \in [0,1]$; (ii) $K(x_i, x_j) = K(x_j, x_i)$; and (iii) $K(x_i, x_i) = 1$. Since the properties of reflexivity and symmetry are satisfied, Exponential kernel induces the fuzzy relation and is denoted by R_E^n. Assume $x = (x_1, x_2, \ldots, x_n) \in R^n$, the Exponential kernel $R_E^{(k)} = \exp\left(-\frac{\|x_i - x_j\|}{\theta}\right)$ is the similarity of samples x_i and x_j with respect to attribute k, and θ is the kernel parameter. Exponential kernel functions can be expressed by a T-norm based combination of reflexive functions. Let $x = (x_1, x_2, \ldots, x_n) \in R^n$ and $y = (y_1, y_2, \ldots, y_n) \in R^n$, Exponential kernel $\exp\left(-\frac{\|x-y\|}{\theta}\right) = \prod_i^n \exp\left(-\frac{(x_i - y_i)}{\theta}\right)$, its product is a T-norm. To measure uncertainty involved in the Exponential kernel approximation, the fuzzy information entropy has been used [15], [16].

Let $I = (U, A)$ be an information system, $A = C \cup D$ be the condition and decision attribute, R_E is fuzzy T-equivalence relation on U computed with Exponential kernel in a sample space $B \subseteq C$. U is divided into $\{d_1, d_2, \ldots, d_l\}$ with the decision attribute. The fuzzy positive region D contain all objects of U that can be classified into classes of U/D using the information available in B is given by:

$$POS_B(D) = \cup_{x \in d_i} \underline{R_E} d_i \,, i = 1, 2, \ldots, l.$$

One of the most important issues of rough sets in data analysis is discovering dependencies between attributes. Jensen *et al.*, [17] generalized the function of dependency in the case of fuzzy sets and proposed a fuzzy dependency function. If a set of attribute D completely depends on attribute B, then $B \Rightarrow D$. Dependency can be defined in a concise way by using rough sets, for any $B, D \subset A$, D depends on B in a degree k, denoted by $B \Rightarrow_k D$, if $k = \gamma_B(D) = \frac{|POS_B(D)|}{|U|} = \frac{\left|\cup_{x \in d_i} \underline{R_E} d_i\right|}{|U|}$, where $0 \le k \le 1$. If $k = 1$, D depends on B completely. Given a decision table $\langle U, A, V, f \rangle$, if $a \in B$, $B \subseteq C$ and $A = C \cup D$, then the condition attribute a is indispensible if

$\gamma_{\{B-a\}}(D) < \gamma_B(D)$, otherwise a is redundant. If $B \subseteq C$ is independent if any $a \in B$ is indispensable. The reduction of attributes can be performed by comparing equivalence relations generated by sets of attributes. Attributes are removed so that the reduced set provides the same quality of classification as the original. A reduct is defined as a subset of minimal cardinality R_{min} of the conditional attribute C such that $\gamma_R(D) = \gamma_C(D)$ [17], [18]. Also the significance of a in B as $Sig(a, B, D) = \gamma_B(D) - \gamma_{B-a}(D)$[19]. The significance of attribute is co-related with three variables: a, B and D. The attributes significance is different for each decision attribute if they are multiple decision attributes in a decision table. This is applicable to backward feature selection algorithm, where the redundant features are evaluated from the original features one by one. Again, for forward feature selection, the significance of attribute is $Sig(a, B, D) = \gamma_{B \cup a}(D) - \gamma_B(D)$, $\forall \in A - B$.

4. FEATURE SELECTION WITH EXPONENTIAL KERNEL BASED FUZZY ROUGH SETS

Feature selection algorithm finds the dependency relations between the attribute sets to find more efficient representation of the data where each sample belongs to its decision with the highest certainty and there is no redundant attribute in feature subspace. The measure of fuzzy dependency is outlined in Algorithm1. This algorithm is to iteratively estimate feature weights according to their ability to discriminate between neighboring patterns. In each iteration, a pattern x is randomly selected and then two nearest neighbors of x are found, one from the same class (M) and the other from the different class (N). The average weight of the i^{th} feature is then updated.

Algorithm1. Dependency with Exponential kernel approximation

Input: $\langle U, B, D \rangle$ and parameter θ
// U is the set of samples;
// B is the feature set;
// D is the set of all decision attributes.
// Output: dependency $\gamma_B(D)$
Step 1: $\gamma_B(D) = 0$
Step 2: for $i = 1$ to m
Step 3: find the nearest samples M_i (of same class) and N_i (of different class) of objects x_i
Step 4: $\gamma_B(D) = \gamma_B(D) + \sqrt{1 - \left(\exp\left(-\frac{\|x_i - M_i\| + \|x_i - N_i\|}{2\theta}\right)\right)}$
Step 6: $\gamma_B(D)$
Step 5: End

Algorithm1 finds the significance of features and ranks the features. The evaluation of the dependency function performed for a dataset containing m attributes and n objects has time complexity $O(mn \log n)$, which is same as that of Relief [20]. Redundant features are computed from the original set of features one by one.

Feature subset selection is outlined in Algorithm2. In each iteration the algorithm begins with an empty set R of attribute and adds one feature, which makes the increment of dependency, into the set R. For each iteration a conditional feature that has not already been evaluated will be temporarily added to the subset R. The subset is then evaluated by maximizing the increment of dependency. The algorithm continues to evaluate the subsets until the dependency of the current reduct candidate equals to zero by adding any new feature into the attribute subset R.

Algorithm2: Feature selection

Input: $\langle U, C, D \rangle$ and threshold ε
// U is the set of samples;
// C is the set of all conditional attributes;
// D is the set of all decision attributes.
// Output: a reduction R
Step 1: $\emptyset \rightarrow R, \gamma = 0$
Step 2: For each $x \in (C - R)$
Step 3: Compute $\gamma_i = \gamma_{\{x\} \cup R}(D)$
Step 4: Compute $Sig(x, R, D) = \gamma_{\{x\} \cup R}(D) - \gamma_R(D)$
Step 5: Select the attribute x and find max $(Sig(x, R, D))$
Step 6: if $\gamma_i - \gamma_R(D) > \varepsilon$, // ε is a small positive number to control the convergence
Step 7: $R \cup \{x\} \rightarrow R$
Step 8: $\gamma_i \rightarrow \gamma_R$
Step 9: else return R
Step 10: End

The feature selection algorithm is based on forward search strategy, the first step is the increment of dependence of the current candidate subset which maximizes each selected attribute in time complexity $O(mn \log n)$ and the second step is to analyze whether the sample is consistent with time complexity $O(m)$. So the worst case of computational complexity is $O(m^2 n \log n)$, which is same as in [21], where m and n are the numbers of attributes and objects.

5. EXPERIMENTATION

This section presents the results of experimental results obtained for six real-valued datasets. The datasets have been downloaded from UCI repository of machine learning databases [22]. The data sets used in the present work are outlined in Table 1.

Table 1: Dataset description

Sl. No.	Dataset	Samples	Features	Class
1	Dermatology (derm)	366	33	6
2	Hepatitis (hepa)	155	19	2
3	Ionosphere (iono)	351	34	2
4	Wisconsin diagnostic breast cancer (wdbc)	569	31	2
5	Wisconsin prognostic breast cancer (wpbc)	198	33	2
6	Wine recognition (Wine)	178	13	3

Learning algorithms *viz.* Classification and Regression Tree (CART) [23] and linear Support Vector Machine (SVM) [2] are used to evaluate the selected features. The results are obtained with 10-fold cross validation mode. The parameters of SVM are taken as the default values (the value of kernel type is 0, kernel parameter is 0.05 and cost factor is 1) [24]. To compute the membership grades of samples belonging to the lower approximation of decision with Exponential kernel experimentation is carried out with different values of kernel parameter (θ) over different datasets and the dependency of decision to each feature set is obtained [14], [25], [26]. This dependency becomes a good estimate of the classification abilities of the corresponding features. Higher values of θ are

reflective of the classification capabilities of the respective features. The dependency goes up firstly and gets some peak, and then decreases. For the evaluation of a feature, the optimal interval for the values of the kernel parameter θ is taken as [0.01, 0.5].

The performance of the Exponential kernel based fuzzy rough sets for feature selection is compared with Gaussian kernel approximation [27] and the neighborhood rough sets (NRS) [21], [28]. The evaluating function reflects the classification performance in feature selection and feature ranking.

The features selected with different models based on the significance value for the derm, hepa, iono, wdbc, wpbc and wine data sets are tabulated in Table 2 and the number of features selected by different models is tabulated in Table 3. The order of the features depicted in the Table 2 is the orders in which the features are being added to the feature space. In feature selection, the first best features are selected in ranking and best feature are added one by one and the classification performance of the current features in each round is determined until the classification performance does not improve significantly when adding more features.

From Table 2 it is seen that the significance value increases with the corresponding number of selected features. The significance value does not increase once it reaches its maximum value. Features that have been added during the learning process are the selected features. From Table 3 it is observed that Exponential kernel based attribute reduction algorithm gives better reduction rate for the datasets considered in the experimentation. Three best results of high reduction rate are shown in Figs. 1-3.

After performing reduction the dependency of reduced features is checked. The dependency of Exponential kernel with different kernel parameter values (θ) for both original features and reduced features based on Exponential fuzzy rough set (for different datasets) is shown in Table 4.

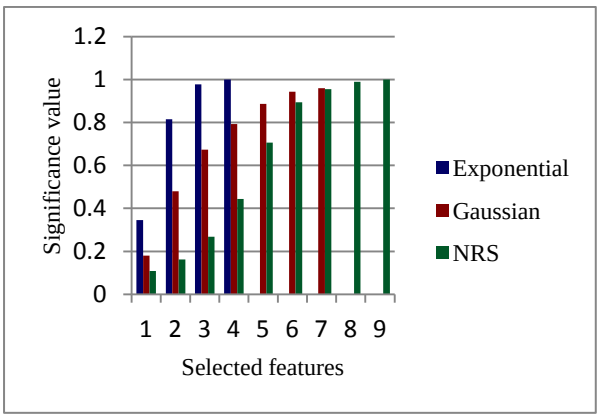

Fig.1. Significance value of selected features by using Exponential, Gaussian and Neighborhood rough sets for ionosphere dataset

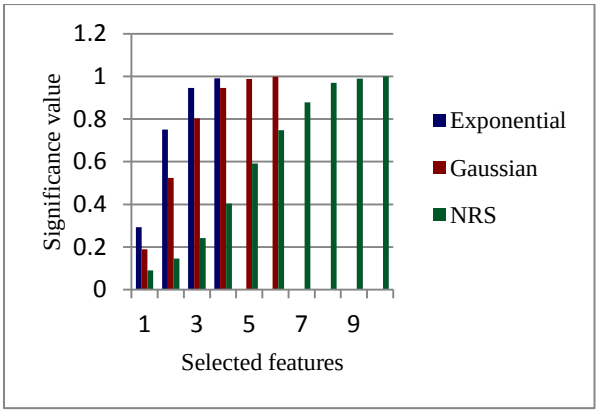

Fig.2. Significance value of selected features by using Exponential, Gaussian and Neighborhood rough sets for wpbc dataset

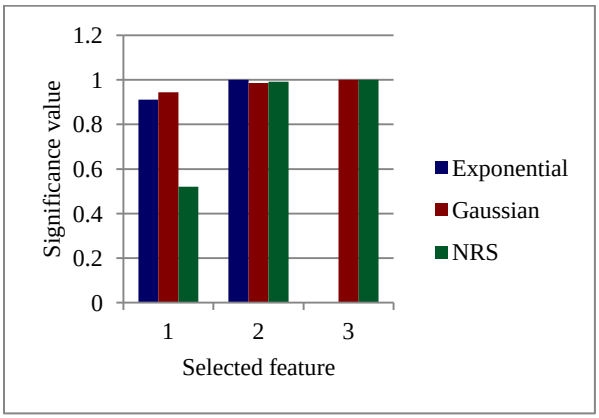

Fig.3. Significance value of selected features by Exponential, Gaussian and NRS feature selection for wdbc dataset

Table 2: Subsets of features selected by feature selection models

Dataset	Exponential kernel approximation		Gaussian kernel approximation		Neighborhood rough sets	
	Selected attribute	Significance value	Selected attribute	Significance value	Selected attribute	Significance value
derm	22	0.2951	22	0.2486	22	0.2486
	33	0.4918	27	0.4454	27	0.4454
	34	0.6667	5	0.5464	5	0.5464
	16	0.8607	15	0.6913	15	0.6913
	1	0.9262	34	0.9016	4	0.7568
	14	0.9672	14	0.9754	31	0.8770
			3	1.0000	16	0.9208
					32	0.9754
					1	0.9891
					8	1.0000
hepa	15	0.0733	17	0.0241	17	0.0194
	16	0.4314	16	0.1707	16	0.1548
	2	0.7625	2	0.5088	2	0.5935
	18	0.8866	18	0.7062	18	0.8000
	17	0.9377	15	0.8024	15	0.9032
	4	0.9699	4	0.8696	4	0.9613
	6	0.9850	6	0.9152	5	0.9871
			7	0.9397	1	1.0000
iono	5	0.3456	5	0.1795	1	0.1083
	6	0.8143	8	0.4794	5	0.1624
	25	0.9774	27	0.6730	13	0.3504
	32	0.9998	24	0.7931	34	0.6410
			33	0.8863	24	0.9145
			34	0.9434	9	0.9915
			22	0.9598	3	1.0000
wdbc	24	0.9112	24	0.9436	24	0.5202
	4	1.0000	4	0.9862	4	0.9912
			22	0.9997	2	1.0000
wpbc	2	0.2923	2	0.1897	2	0.0909
	13	0.7500	13	0.5234	30	0.1465
	29	0.9454	33	0.8030	14	0.2424
	7	0.9908	24	0.9459	3	0.4040
			25	0.9874	33	0.5909
			11	0.9988	13	0.7475
					4	0.8788
					11	0.9697
					7	0.9899
					27	1.0000
wine	10	0.3592	10	0.2215	10	0.0169
	7	0.8806	13	0.7281	13	0.2528
	13	0.9818	7	0.9385	7	0.4944
	11	0.9990	11	0.9873	11	0.7191
			5	0.9973	8	0.8539
					12	0.9382
					5	0.9888
					1	1.0000

Table 3: Number of features selected (entries in bold are the best cases)

Sl. No.	Dataset	Raw data	Exponential kernel	Gaussian kernel	Neighborhood rough sets
1	derm	33	**6**	7	10
2	hepa	19	**7**	8	8
3	iono	34	**4**	7	9
4	wdbc	31	**2**	3	3
5	wpbc	33	**4**	6	10
6	wine	13	**3**	5	**3**

Table 4: Dependency of Exponential kernel with different parameter (θ)

Dataset	Parameter (θ)	Dependency with Exponential kernel		Dataset	Parameter (θ)	Dependency with Exponential kernel	
		Original dataset	Reduced dataset			Original dataset	Reduced dataset
derm	0.01	1.000	1.000	wdbc	0.01	1.000	1.000
	0.02	1.000	1.000		0.02	1.000	1.000
	0.05	1.000	1.000		0.05	1.000	1.000
	0.08	1.000	1.000		0.08	1.000	1.000
	0.1	1.000	1.000		0.1	1.000	1.000
	0.3	1.000	1.000		0.3	1.000	1.000
	0.5	0.9980	0.9976		0.5	0.9999	0.9991
	0.8	0.9674	0.9957		0.8	0.9990	0.9735
hepa	0.01	1.000	0.5032	wpbc	0.01	1.000	1.000
	0.02	0.9996	0.5032		0.02	1.000	1.000
	0.05	0.9929	0.5032		0.05	1.000	1.000
	0.08	0.9685	0.5032		0.08	1.000	0.9908
	0.1	0.9463	0.5032		0.1	0.9988	0.9445
	0.3	0.6945	0.5032		0.3	0.9837	0.8984
	0.5	0.5212	0.5010		0.5	0.8859	0.5122
	0.8	0.3667	0.4716		0.8	0.7023	0.3378
iono	0.01	0.9943	1.000	wine	0.01	1.000	1.000
	0.02	0.9943	1.000		0.02	1.000	1.000
	0.05	0.9930	0.9998		0.05	1.000	1.000
	0.08	0.9897	0.9964		0.08	1.000	1.000
	0.1	0.9862	0.9910		0.1	1.000	1.000
	0.3	0.9087	0.8482		0.3	0.9537	1.000
	0.5	0.8128	0.6753		0.5	0.7866	0.9983
	0.8	0.6937	0.4955		0.8	0.5711	0.9766

From Table 4 it is observed that for most of the cases Algorithm2 gets the good sub set of features for classification with kernel parameter values (θ) are in the interval [0.1, 0.5]. From experimental results tabulated in Table 2 the best results are obtained for the wdbc, wpbc and iono dataset. It is seen that feature set {24,4} are the best features and it ranks first and second for wdbc dataset in all the methods and feature 2 is the best single feature for wpbc dataset and it ranks first in all the methods. For the iono dataset, the order of features induced by Exponential kernel approximation is 5, 6, 25, 32; while the order of features induced by neighborhood rough sets is 1, 5, 13, 34, 24, 9, 3. Features 5 is the best feature, it ranks first with Exponential kernel approximation and the Gaussian kernel approximation while it ranks second with the neighborhood rough sets. Attribute reduction algorithms are sensitive to these differences. It is noted here that these little differences may leads to completely different feature subsets in feature selection algorithms.

The classification accuracy of the corresponding selected features is then evaluated with the learning algorithms, *viz.* CART and SVM, to test the quality of the selected subsets of features. The classification performances for the raw data and the reduced data based on 10-fold cross validation are shown in Table 5 where the values in bold shows the highest accuracy with the reduced datasets. The CART with Exponential kernel approximation outperforms the other approaches *viz.* Gaussian kernel approximation and NRS for the ionosphere and wine datasets. With regard to SVM, Exponential kernel approximation outperforms the Gaussian kernel approximation and NRS for the derm, iono, wdbc, wpbc and wine datasets. It is noted that Exponential kernel approximation is little weaker than Gaussian and neighborhood rough sets with the CART learning algorithm. However, Exponential kernel approximation produces the best performances over the Gaussian kernel and NRS feature selection methods with the SVM learning algorithm.

Table 5: Classification accuracy based on CART and SVM

Dataset	Classifier	Classification Accuracy			
	CART/SVM	Before feature selection	Exponential kernel	Gaussian kernel	NRS
derm	**CART**	0.9226	0.9375	0.9200	**0.9970**
	SVM	0.8797	**0.9945**	0.9890	0.9882
hepa	**CART**	0.8249	0.8309	**0.8375**	0.7333
	SVM	0.8717	0.8082	0.7808	0.8000
iono	**CART**	0.8755	**0.8947**	0.8922	0.8940
	SVM	0.9117	**0.9185**	0.9136	0.8834
wdbc	**CART**	0.9050	0.9086	0.9069	0.9244
	SVM	0.9462	**0.9550**	0.9154	0.9347
wpbc	**CART**	0.7121	0.6847	0.7153	0.7132
	SVM	0.6000	**0.7692**	0.6363	0.7142
wine	**CART**	0.8694	**0.9222**	0.9222	0.9056
	SVM	0.9213	**0.9550**	0.9250	0.9375

From the results obtained it is seen that most of the features in all data sets are deleted and all the algorithms produce distinct subset of features. Also all the algorithms do not get the same subset of features for any data set in the experiments. As different learning algorithms make use of available features in distinct ways, different learning algorithms may require different feature subsets to produce the best classification performance. It is easy to find from Table 5 that CART obtains higher classification accuracy with regard to the iono and wine datasets whereas SVM obtains higher classification accuracy with regard to the derm, iono, wdbc, wpbc and wine datasets when Exponential kernel based fuzzy rough set model is used for feature selection as compared to other feature selection methods *viz*. Gaussian kernel and NRS.

From the analysis of the results it is seen that Exponential kernel approximation based attribute reduction algorithm gives better reduction rate for the datasets considered in the experimentation as compared to two other attribute reduction algorithms. At the same time, the reduced data improves the classification performance of the raw datasets.

6. CONCLUSION

In this paper an exponential kernel function is integrated with fuzzy rough set to develop the Exponential kernelized fuzzy rough set for feature selection. To demonstrate the effectiveness of the Exponential kernel based fuzzy rough set for feature selection experimentation is carried out over six benchmark data sets obtained from the public domain repository. The experimental results obtained with the considered data sets show that the Exponential kernel based fuzzy rough set model for feature selection is effective and improves the classification accuracy in comparison to Gaussian kernel approximation and NRS for the ionosphere and wine data sets with CART learning algorithm. The Exponential kernel based fuzzy rough set model for feature selection also outperforms the Gaussian kernel and NRS and improves the classification accuracy for the derm, iono, wdbc, wpbc and wine data sets with SVM learning algorithm. Further investigation may include the study of the relationship between the statistical property of kernel methods and the class imbalance problem of datasets for the task of feature selection.

7. REFERENCES

[1] Vapnik, V. 1995. The Nature of Statistical Learning Theory. Springer, New York.

[2] Cristianini, N., Shawe-Taylor, J. 2000. An Introduction to Support Vector Machines and other Kernel-based Learning Methods. Cambridge University Press, Cambridge.

[3] Shawe-Taylor, J., Cristianini, N. 2004. Kernel Methods for Pattern Analysis. Cambridge University Press.

[4] Pawlak, Z. 1982. Rough sets, Int. J. Inform. Comput. Sci. 11: 314–356.

[5] Pawlak, Z. 1991. Rough Sets – Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht.

[6] Dubois, D. and Prade, H. 1990. Rough fuzzy sets and fuzzy rough sets, Int. J. General Syst. 17 (2–3) : 191–209.

[7] Wu,W. and Zhang, W. 2004. Constructive and axiomatic approaches of fuzzy approximation operators, Inf. Sci. 159 (3–4) : 233–254.

[8] Yeung, D.S., Chen, D., Tsang, E.C.C., Lee, J.W.T. and Wang, X.Z. 2005. On the generalization of fuzzy rough sets, IEEE Trans. On Fuzzy Systems 13 (3) 343–361.

[9] Moser, B. 2006. On the t-transitivity of kernels, Fuzzy Sets Syst. 157 :787–1796.

[10] Moser, B. 2006. On representing and generating kernels by fuzzy equivalence relations, J. Mach. Learn. Res. 7 : 2603–2620.

[11] Morsi , N. N. and Yakout,M. M. 1998. Axiomatics for fuzzy rough set, Fuzzy Sets Syst. 100 : 327–342.

[12] D. Dubois and H. Prade, 1992. Putting rough sets and fuzzy sets together, Intelligent Decision Support, Kluwer Academic Publishers, Dordrecht, 203–232.

[13] P. Lingras and R.Jensen. 2007. "Survey of Rough and Fuzzy Hybridization", IEEE Intl. Conf. on Fuzzy Systems, 1-6.

[14] Genton, M. 2001. Classes of kernels for machine learning: A statistics perspective, Journal of Machine Learning Research, 2: 299–312.

[15] Hu, Q.H., Yu, D.R. and Xie, Z.X. 2006. Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognition Lett. 27 (5) : 414–423.

[16] Yu, D., Hu, Q.H. and Wu, C. 2007. Uncertainty measures for fuzzy relations and their applications, Appl. Soft Comput. 7 : 1135–1143.

[17] Jensen, R. and Shen, Q. 2004. Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough-Based Approaches, IEEE Trans. On Know. and Data Engg. 16(12) : 1457–1471.

[18] Jensen, R. and Shen, Q. 2009. New approaches to fuzzy-rough feature selection, IEEE Trans. Fuzzy Syst. 17: 824–828.

[19] Hu, Q., Yu, D. and Xie, Z. 2008. Neighborhood classifiers, Expert Syst. Appl. 34 : 866–876.

[20] Robnik-Sikonja, M. and Kononenko, I. 2003. Theoretical and empirical analysis of ReliefF and RReliefF, Mach. Learning 53 : 23–69.

[21] Hu, Q., Yu, D., Liu, J. and Wu, C. 2008. Neighborhood rough set based heterogeneous feature subset selection. Inf. Sciences 178 3577–3594.

[22] Blake ,C. Merz, C., Hettich, S. and Newman, D. J. 1998. UCI repository of machine learning databases, University of California, School of Information and Computer Sciences, Irvine, CA.

[23] Brieman, L., Friedman, J., Stone, C.J. and Olshen, R.A. 1984. Classi_cation and Regression Trees, Chapman and Hill.

[24] Yan, R. 2006. A MATLAB Package for Classification Algorithm.

[25] Sun, Y. 2006. Iterative RELIEF for feature weighting: algorithms, theories and applications, IEEE Trans. Pattern Analysis and Machine Intelligence 1-27.

[26] Atkeson, C. G. Moore, A. W. and Schaal, S. 1997. Locally weighted learning, Artificial Intelligence Review, 11(15) : 11-73.

[27] Hu, Q., Zhang, L., Chen, D., Pedrycz, W. and Yu, D. 2010. Gaussian kernel based fuzzy rough sets: Model, uncertainty measures and applications, Intl. Journl. of Approx. Reasoning 51 : 453-471.

[28] Lin, T.Y. 2001. Granulation and nearest neighborhoods: rough set approach. In: Pedrycz, W. (ed.) Granular computing: an emerging paradigm, pp. 125–142, Physica-Verlag, Heidelberg.