Journal of Network and Innovative Computing
ISSN 2160-2174 Volume 2 (2014) pp. 178-186
© MIR Labs, www.mirlabs.net/jnic/index.html

Hill-climber Based Fuzzy-Rough Feature Extraction with an Application to Cancer Classification

Sujata Dash

Department of Computer Application,
North Orissa University, Baripada, Odisha, India.
Sujata_dash@yahoo.com

Abstract: **Real-world problems are often imprecise and redundant thereby create difficulty in taking decisions accurately. In recent past, rough set theory has been used for predicting potential genes responsible for causing cancer using discrete dataset. But discretization of data makes the dataset inconsistent by loosing information. To overcome this problem, this paper presents an efficient approach to predict the dominant genes using fuzzy-rough boundary region-based feature selection in combination with a heuristic hill-climber search method. But hill-climber search method produces subsets that contain redundant features. This problem is addressed using fuzzy-rough boundary region-based method that finds the reduct by minimizing the total uncertainty degree of the dataset to achieve faster convergence. Hill-climber based fuzzy-rough boundary region generates fuzzy decision reducts, which represent the minimal set of non-redundant features, capable of discerning between all objects. In this work, we attempt to introduce a prediction scheme that combines the proposed filter method with three different rule classifiers such as JRIP, Decision Tree and PART. We demonstrate the performance of the model by two benchmark microarray datasets and the results show that our proposed method significantly reduces the dimensionality while preserving the classification accuracy. The function of selected genes are classified and validated from gene ontology website, DAVID, which shows the relationship of genes with the disease.**

Keywords: *Fuzzy set, rough set, fuzzy-rough set, hill-climber search, feature extraction.*

I. Introduction

The development of microarray technology helps us to access DNA microarray datasets containing millions of expression level of genes. Recognizing classes of cancer based on gene expression levels is important for cancer diagnosis [1]. Feature extraction methods are used to extract/select the most predictive genes/features preserving the original meaning of the features after reduction. These predictive genes are responsible for causing cancer and become a key issue for cancer diagnosis.

The existing methods of attribute reduction models reduce the genes of microarray dataset can be broadly classified into three classes, such as; filter, wrapper and embedded method. In filter method, features are reduced based on the individual characteristics of the attribute, which determines their relevance or prediction power with regard to the target classes. In wrapper method, feature selection is “wrapped”' around a learning method and the effectiveness of a feature is directly judged by the estimated accuracy of the learning method. In general, a high-dimensional dataset increases the chances of finding redundant patterns by the classifier, which are not valid. However, feature reduction is an indispensable component [2], [3], [4] of real life classification problems.

Fuzzy sets [5] and rough sets [6] address two complementary characteristics of imprecise data and knowledge: the former model expresses vague information of the samples belong to a relation to a given degree, while the later provide approximations of concepts in the presence of incomplete information. The following three aspects has made rough-set theory a successful method in feature reduction : 1) Only the facts hidden in data are analyzed, 2) No additional information about the data is required for data analysis such as thresholds or expert knowledge on a particular domain, 3) It finds a minimal knowledge representation of data. Of late researchers have taken interest in developing efficient hybrid methodologies [7] which are capable of dealing with imprecision and uncertainty of the problem. Moreover, such developments offer a high degree of flexibility and provide robust solutions and advanced mechanisms for data analysis. The hybrid fuzzy- rough sets encapsulate the related but distinct concepts of vagueness of fuzzy sets [8] and indiscernibility of rough sets. Both of them are complementary to each other and can be encountered in real-life problems. A fuzzy-rough set is an approximation of a crisp set or a fuzzy set in a fuzzy approximation space. The fuzzy-rough set model may be used to reveal the knowledge hidden in fuzzy decision systems. Therefore, fuzzy-rough sets have the advantages of rough sets while reducing the information lost in real-valued data sets caused by the discretization in rough sets [9]. But with respect to the complexity of fuzzy rough sets, the amount of investigation done by researchers on attribute reduction method in fuzzy rough set theory seems not very prevalent and also an expensive solution to the problem. Above all, finding of minimal reducts are NP-hard problem established by Skoworn [10]. Therefore, to select minimal reduct sets, heuristics or stochastic approaches have to be considered.

There are two approaches for reducing features in fuzzy-rough method such as hill-climbing or greedy methods and stochastic methods. The reduced significant attributes of fuzzy-rough set is used as a heuristic knowledge for hill-climbing method. The hill-climbing method starts with an empty set and then employs forward selection or backward elimination. The forward selection adds the most significant attribute one at a time from the candidate set, until the selected set is a reduct. On the other hand, the backward elimination starts with full attribute set and removes the attribute incrementally. X. Hu has used the positive region based attribute significance and Wang has used conditional entropy-based attribute significance as heuristic knowledge to devise a reduction algorithm [11], [12]. The positive region and conditional entropy-based methods select a minimal subset of features that represent the whole dataset. K. Hu [13] has proposed a heuristic reduction algorithm for finding significant attributes, making use of discernibility matrices. This method selects minimal subset of attributes with high discriminatory power and maximal between class separability for the reduct datasets. Shannon’s entropy and its variants have been applied to measure uncertainty in rough set theory from information theory point of view. However, few studies have been done on attribute selection in incomplete decision systems based on information-theoretical measurement. J. Dai et. al [23] has proposed new form of conditional entropy to measure the importance of attributes in incomplete decision systems. Based on this, they have constructed three attribute selection approaches, including an exhaustive search strategy approach, a greedy (heuristic) search strategy approach and a probabilistic search approach for incomplete decision systems. The method is tested with many real life datasets and concluded that two of these methods are effective for are effective for attribute selection in incomplete decision system.

It is also observed that, all approaches to fuzzy-rough and rough set feature selection use lower approximation for the evaluation of feature subsets. The information available in the lower approximation represents the certainty of object membership to a given concept whereas upper approximation represents the degree of uncertainty of objects. Hence the objects within the boundary region will have less uncertainty and will be more useful. Some researchers have also integrated stochastic feature extraction approach with rough set theory. Zhai [14] proposes an integrated feature extraction method based on rough set and genetic algorithms. Xyangyang [15] finds minimal rough reducts using another stochastic strategy, Particle Swarm Optimization. However, it uses highly time-consuming operations and cannot assure that the resulting subset is a reduct set. They found their methods provide an approximated solution at the expense of increased computational effort. Cadenas [24] proposed a feature selection technique which can work with both crisp and low quality imprecise and uncertain data. The technique is based on Fuzzy Random Forest method and integrates filter and wrapper techniques into a sequential search procedure which improves the classification accuracy. This approach consists of the following steps: (1) scaling and discretization process of the feature set; and feature pre-selection using the discretization process (filter); (2) ranking process of the feature pre-selection using the Fuzzy Decision Trees of a Fuzzy Random Forest ensemble; and (3) wrapper feature selection using a Fuzzy Random Forest ensemble based on cross-validation. The efficiency and effectiveness of this approach is proved through several experiments using both high dimensional and low quality datasets. The approach shows a good performance not only in terms of classification accuracy, but also in features selection and good behaviour both with high dimensional datasets (microarray datasets) and with low quality datasets.

In this paper, I have proposed an integrated heuristic reduction algorithm, investigating how an integrated hill-climber based fuzzy-rough boundary region attribute reduction (HCBFRBAR) method can be applied to find minimal reducts. In conventional RSAR, a reduct is a subset R of attributes that have the same information content as the full attribute set A. But, this is not necessarily true in fuzzy-rough approach as the uncertainty encountered when objects belong to many fuzzy equivalence classes results in a reduced total dependency. Using this concept, a fuzzy-rough hill-climbing search algorithm based on fuzzy similarity is developed for locating fuzzy-rough reducts. This algorithm minimizes the total uncertainty degree instead of maximizing the dependency degree. Again, the time complexity of the algorithm is same as that of FRFS but avoids the Cartesian product of fuzzy equivalence classes. It calculates the fuzzy boundary region considering both upper and lower approximations which is more complex than that of the lower approximation alone. Hill-climbing methods are more efficient when dealing with minimum noise and a small number of interacting features. Hill-climbing is best suited to problems where the heuristic gradually improves the closer it gets to the solution. It works poorly where there are sharp drop-offs. It assumes that local improvement will lead to global improvement. In this algorithm, the hill-climbing search method finds optimal regions of complex search space using fuzzy-rough boundary region-based improved attribute as heuristic knowledge. The performance of the proposed integrated model is evaluated using binary and multi-class microarray datasets and also compared with the results of three integrated heuristic algorithms such as hill-climber based fuzzy-rough discernibility matrix (HCBFRDMAR), fuzzy-entropy based attribute reduction (HCBFREBAR) and vaguely quantified fuzzy-rough lower approximation attribute reduction (HCBFRVQLAR). The studies show that it has a strong search capability in the problem space and more effective than conventional rough sets and fuzzy-rough based approaches. The Functional Classification Tool of DAVID Gene Ontology has been applied which provides a rapid means to organize large lists of genes into functionally related groups to help unravel the biological content captured by high throughput technologies.

The rest of the article is organised as follows. Section 2 illustrates fuzzy-rough attribute reduction method focusing on dimensionality reduction. Section 3 describes the proposed hill-climber based fuzzy-rough boundary attribute reduction [HCBFRBAR] algorithm [16] . Section 4 shows experimental results on two benchmark data sets, multi-class Leukemia [17] and binary class Prostrate cancer [18]. Proposed attribute reduction technique is compared with the results of three integrated heuristic fuzzy-rough attribute reduction based on fuzzy similarities and validation of the obtained results presented in section 5. Finally, a conclusion and future scope of research is presented in section 6.

II. Background of Fuzzy-Rough Set

A. Fuzzy-Rough Attribute Reduction

Rough set theory can be used to discover data dependencies and reduction of the attributes contained in the dataset using the data available in the dataset [22]. The main drawback of rough set theory is that it cannot deal with real-valued problems, whereas many real world problems are real-valued. Fuzzy–rough set theory is a mathematical technique which is capable of reducing crisp to real valued attribute datasets. The core of the Fuzzy–Rough Attribute Selection is the concept of indiscernibility relation which partitions the domain. Given a set of attributes as the objects of the domain, objects with the same attribute values are indiscernible and would belong to the same block of the partition. The task is to approximate a *rough* (imprecise) concept in the domain by a pair of *exact* concepts. These exact concepts are called the lower and upper approximations and are determined by the indiscernibility relation. The formal definitions are given as follows:

Definition1. Let I = (U, A) be an information system, U consists of non-empty finite set of objects and A be the non-empty, finite set of attributes or features a, such that a :U → Va, where Va is a value set. We use information systems called decision table, which contains two types of attributes (A = AC ∪ AD) in which AC is condition attribute and AD is decision attribute. In particular, the decision table shall have a single decision attribute and will be consistent for objects x, y for each condition attribute A, if A(x) = A(y), then d(x) = d(y).

Definition2. For any S ⊆A there is an associated equivalence relation IND(S) which is defined as:

$$IND(S) = \{(x, y) \in U^2 \mid \forall a \in S,\ a(x) = a(y)\}. \qquad (1)$$

IND(*S*) partitions universe *U* in to equivalence classes (*U/S*).

$$[x_i]S = \{x_j \in U: (x_i, x_j) \in IND(S)\},\ x_i \in U. \qquad (2)$$

Definition3. If S and O be two equivalence relations over U, then the positive, negative and boundary regions can be defined as:

$$POS_S(O) = \bigcup_{X \in U/O} \underline{S}X,$$
$$NEG_S(O) = U - \bigcup_{X \in U/O} \overline{S}X, \qquad (3)$$
$$BND_S(O) = \bigcup_{X \in U/O} \overline{S}X - \bigcup_{X \in U/O} \underline{S}X.$$

The positive region contains all objects of U that can be classified into classes of U/O using information from attribute S. The negative region contains objects which cannot be classified into classes of U/O. An attribute b ∈S (⊆AC) is O-dispensable in S if $POS_S(O)$ = POS_S \{b}(AD), otherwise b is O-indispensible in S.

Definition4. A set of attributes O, is completely depend on another set of attributes S, denoted: S ⇒O, if all attributes of O are uniquely determined by values of attributes from S. If functional dependencies were detected, then O is totally depending on S. In rough set theory degree of dependency is defined as:

$$k = \gamma_S(O) = \frac{|POS_S(O)|}{|U|} \qquad (4)$$

if k = 1 then O totally depends on S, k = 0 means that O does not depend on S and 0 < k <1 means that O partially depends on S.

Definition5. In a decision table I = (U, AC∪AD), by eliminating redundant condition attributes, reduct is computed.

B. Fuzzy equivalence classes

Gene expression data is a real-valued dataset, thus we employ fuzzy–rough set for attribute selection. Fuzzy equivalence classes are the core of fuzzy–rough set. In this case decision and conditional values should all be fuzzy. Fuzzy S-lower and S-upper approximations are defined as:

$$\mu_{\underline{S}X}(x) = \sup_{F \in \frac{U}{S}} \min(\mu_F(x), \inf_{y \in U} \max\{1 - \mu_F(y), \mu_X(y)\})$$
$$\mu_{\overline{S}X}(x) = \sup_{F \in \frac{U}{S}} \min(\mu_F(x), \inf_{y \in U} \min\{1 - \mu_F(y), \mu_X(y)\}) \qquad (5)$$

where S is an equivalence class, X is the concept to be approximated and F is a fuzzy equivalence class belonging to U/S.

C. Fuzzy–rough reduction process

The membership degree of an observation x of the universe, belonging to fuzzy positive region, can be defined as:

$$\mu_{POS_S}(Q) = \sup_{X \in U/S} \mu_{\underline{S}X}(x). \tag{6}$$

The observation x will belong to the positive region only if its equivalence class does so.

$$\gamma'_S(Q) = \frac{|\mu_{POS_S(Q)}(x)|}{|U|} = \frac{\sum_{x\in U} \mu_{POS_S(Q)}(x)}{|U|} \tag{7}$$

The dependency of O on S is a proportion of observations which are discernible from the whole dataset.

From theoretical viewpoint, the lower approximation contains information regarding the degree of certainty of objects membership to a given concept. It is also observed that the upper approximation provides information about the degree of uncertainty of objects and hence, this subset will be useful for finding objects associated with less uncertainty in the boundary region. The above issue motivates the development of the proposed method in this paper.

III. Proposed Hill-Climber Based Fuzzy-Rough Boundary- region Attribute Reduction Method [HCBFRBAR]

In most of the approaches to crisp rough set feature reduction and all approaches to fuzzy-rough feature reduction, lower approximation has been used for the evaluation of feature subsets. In conventional rough-set attribute reduction, a reduct R is considered as a subset of the attributes which represent the information of the entire attribute set A. In crisp rough set feature reduction, the boundary region will be zero for each concept when a reduct is obtained but in case of fuzzy-rough method the object memberships to the boundary region for each concept decreases until a minimum is achieved in the search process of optimal subset. Due to the uncertainty involved in fuzzy-rough feature reduction method, the boundary region will not be zero for each concept. The uncertainty involved in each concept X using features in S can be calculated as:

$$U_S(X) = \frac{\sum_{x\in U} \mu_{BND_S(X)}(x)}{|U|} \tag{8}$$

This shows the extent to which the objects belong to the fuzzy boundary region for the given concept X. Then for all the concepts, the total uncertainty for a given feature subset S, is calculated as:

$$\lambda_S(Q) = \frac{\sum_{X\in U/Q} U_S(X)}{|U/Q|}. \tag{9}$$

The above derivation is similar to conditional entropy measure which uses a combination of conditional probabilities H(Q|S) to measure the uncertainty associated with the features in S. In crisp rough set, the minimization of conditional entropy measure is used to find reducts i.e., if the entropy for a feature subset S is zero, then the subset is a reduct.

Considering the above issues, a heuristic search technique such as hill-climber is integrated to fuzzy-rough boundary region attribute reduction method to develop a QuickReduct type algorithm for finding minimal fuzzy-rough reducts. This algorithm minimizes the total uncertainty associated with the features of the given dataset instead of maximizing the dependency degree. The fuzzy-rough reduct will be achieved when the algorithm reaches the minimum for the given dataset. Based on the above concept, a detailed HCBFRBAR (Hill-Climber Based Fuzzy-Rough Boundary-region Attribute Reduction) method is devised.

Algorithm: HCBFRBAR ($A_{C,} A_D$)

Hill-Climber heuristic search performs a greedy forward or backward search through the space of attribute subsets. It may start with no/all attributes or from an arbitrary point in the space. The search operation will stop when the addition/deletion of any remaining attributes results in a decrease in evaluation.

$A_C = \{g_1, g_2, \ldots, g_k\}$, the set of all genes as conditional features.

$A_D = \{d\}$, the set of decision features corresponds to class label of each sample.

Each attribute g_i is represented by a vector $g_i = \{x_{1,i}, x_{2,i}, \ldots, x_{m,i}\}$, i = 1, 2,….., n , where $x_{k,i}$ is the expression level of gene i at sample k, k = 1,2, …, m.

This algorithm computes the degree of uncertainty of the attributes available in boundary region

```
Step1. Let $R = \{ \}$;
Step2. $\lambda'_{Prev} = 0$; $\lambda'_{min} = 0$;
Step3. do
Step4.     $T \leftarrow R$ ;
// For every attribute $A^j \in A_C - R$, compute the smallest uncertainty of conditional attribute //
Step5.     $U_{A^j\ U\{gi\}}(A_D) \le U_{A^j}(A_D)$
// Select the attribute with smallest total uncertainty then record it //
Step6.       if $\lambda_{A^j U\{gi\}}(A_D) \le \lambda_{A^j}(A_D)$
Step7.                 $T \leftarrow R\, U \{gi\}$ ;
Step8.   $\lambda'_{Prev} \leftarrow \lambda'_{A^j \cup \{gi\}}$ ;
Step9.  $R \leftarrow T$;
Step10. until $\lambda'_{Prev} == \lambda'_{min}$ ;
Step11. Return R.
```

IV. Experimental Results

A. The Datasets

Two public microarray datasets were used to assess the performance of the proposed reduction algorithm. The following is a brief description of these datasets.

Prostate Cancer: This dataset consists of 102 samples. The training set contains 52 prostate tumor samples and 50 non-tumor labelled as normal prostate samples with around 12600 genes which is taken from http://www-genome.wi.mit.edu/mpr/prostate . More information about the raw data is available in [18].

Multi-class Leukemia Cancer: Consists of samples from three different types of acute leukemia, acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML) and MLL.

The training data set has 57 leukemia samples (20 ALL, 17 AML and 20 MLL). Each sample has expression patterns of 12582 genes measured by the Affymetrix oligonucleotide microarray.The test data set consists of 15 samples (4 ALL, 8 AML and 3 MLL). Raw data is available in http://www-genome.wi.mit.edu/cgi-in/cancer/datasets.cgi and more information about the cancer dataset can be obtained from [17].

B. Experimental Setup

The WEKA is a well-known package of data mining tools which provides a variety of known, well maintained classification and filter algorithms. This allows us to do experiments with several kinds of classifiers quickly and easily. The tool is used to perform benchmark experiment. Three classifier learners were employed for the classification of the data namely, Decision Table, JRIP and PART. For the hill-climber fuzzy-rough boundary based attribute reduction method, the Lukasiewicz fuzzy connective with the similarity measure max(min((a(y)-(a(x)-sigma_a)) / (a(x)-(a(x)-sigma_a)),((a(x)+sigma_a)-a(y)) / ((a(x)+sigma_a)-a(x)) , 0) and hill climber as search method is used. After attribute reduction, the datasets are reduced according to the discovered reducts. The reduced datasets are then classified using three decision rule-based classifiers. The classifiers JRip [19], PART [20] and Decision Table [21] were employed for the purpose of evaluating the resulting subsets from the attribute reduction. JRip learns propositional rules by repeatedly growing rules and pruning them. Features are added greedily during growing phase until a termination condition is satisfied. Features are then pruned in the next phase using a pruning metric. Once the rule set is generated, a further optimization is performed. The classification rules are again evaluated and deleted based on their performance on randomized data. In Decision Table, the evaluation measure is used to evaluate the performance of attribute combinations used and a search method is used to find good attribute combinations for building and using a simple decision table majority classifier. PART generates rules by means of repeatedly creating partial decision trees from the data. The algorithms employ a divide –and-conquer strategy such that it removes instances covered by current ruleset during processing. The classification rule is created by building a pruned tree for the current set of instances. The leaf with the highest coverage is considered as a rule.

V. Result Analysis

Table1shows the number of features, instances, categories of sample and the number of features selected using 4 different hill-climber based fuzzy-rough attribute reduction algorithms. It can be observed from the table that except the proposed boundary region based fuzzy-rough attribute reduction algorithm (HCBFRBAR), the other 3 reduction algorithms i.e., hill-climber based fuzzy-rough discernibility matrix (HCBFRDMAR), hill-climber based fuzzy-rough entropy based attribute reduction (HCBFREBAR) and hill-climber based fuzzy-rough vaguely quantified lower approximation attribute reduction (HCBFRVQLAR) select equal number of attributes (4 no’s each) in the reduction process for both the datasets. But fuzzy-rough boundary region based hill-climber attribute reduction (HCBFRBAR) algorithm selects only 3features for multi-class Leukemia dataset and 4 features for the binary-class Prostate cancer dataset.

Table 1. Microarray Datasets and Reduct sizes found from Feature Selection Algorithms

Data sets	**Features**	**Instances**	**Fuzzy Rough Set Reduction Algorithm**			
		Train /Test	***HCBDMAR***	***HCBEBAR***	***HCBVQLAR***	***HCBBAR***
Leukemia	12582	62(57:15)	4	4	4	3
Prostate	12600	102(50:52)	4	4	4	4

Table2. % of Classification Error of Leukemia and Prostate Dataset without using Feature Selection methods with 10 fold-CV.

DT Algorithm, No of selected rules-06		**JRIP , No of selected rules-04**	**PART, No of selected rules-03**
Leukemia	***False Positive Rate***	***False Positive Rate***	***False Positive Rate***
ALL	13.5	13.5	8.1
MLL	12.5	7.5	2.5
AML	8.1	16.2	21.6
Overall Classification	22.8	24.56	21.5

Table3. % of Classification Error of Leukemia and Prostate Dataset without using Feature Selection methods with 10 fold-CV

DT Algorithm, No of selected rules-09		**JRIP , No of selected rules-03**	**PART, No of selected rules-03**
Prostate	***False Positive Rate***	***False Positive Rate***	***False Positive Rate***
Tumor	2.6	1.6	1.2
Normal	5.8	19.2	13.5

Table4. Selected Genes of Leukemia using Fuzzy Rough Set Reduction Algorithms.

HCBFRDMAR	39649_at	32715_at	38017_at	109_at
HCBFREBAR	39318_at	35985_at	650_s_at	109_at

HCBFRVQLAR	39318_at	39155_at	306_s_at	109_at
HCBFRBAR	34833_at	35260_at	102_at	-

Table5. Selected Genes of Prostate Dataset using Fuzzy Rough Set Reduction Algorithms

HCBFRDMAR	41221_at	37720_at	1052_s_at	111_at
HCBFREBAR	37639_at	37707_i_at	38378_at	107_at
HCBFRVQLAR	34791_at	37716_at	39853_at	32598_at
HCBFRBAR	37639_at	40508_at	38408_at	108_g_at

Table 6. Classification Error (%) of Leukemia Dataset using different Reducts with10 – fold CV

Leukemia Dataset		**ALL**	**MLL**	**AML**	**Overall classification**	**No of rules**
HCBFRDM	JRIP	21.6	20.0	13.5	36.8421	3
	PART	16.2	25.0	5.4	31.5789	6
	DT	40.5	10.0	5.4	36.8421	4
HCBFREB	JRIP	5.4	7.5	10.8	15.7895	3
	PART	10.8	15.0	5.4	21.0526	5
	DT	5.4	12.5	10.8	19.2982	3
HCBFRVQ	JRIP	13.5	5.0	10.8	19.2982	3
	PART	5.4	1.0	10.8	17.5439	3
	DT	2.7	7.5	13.5	15.7895	3
HCBFRB	JRIP	5.4	2.5	0.0	5.2632	3
	PART	2.7	5.0	2.7	7.0175	3
	DT	8.1	5.0	2.7	10.5263	6

Table7. Classification Error (%) of Prostate Dataset using different Reducts with 10 – fold CV

Prostate Dataset		**Tumor**	**Normal**	**Overall classification**	**No of rules**
HCBFRDM	JRIP	12.0	25.0	18.6275	3
	PART	6.0	25.0	15.6863	4
	DT	8.0	26.9	17.6471	6
HCBFREB	JRIP	12.0	11.5	11.7647	4
	PART	16.0	11.5	13.7255	4
	DT	18.0	9.6	13.7255	8
HCBFRVQ	JRIP	1.2	3.8	7.8431	3
	PART	1.0	11.5	10.7843	4
	DT	2.0	5.8	12.7451	8
HCBFRB	JRIP	12.0	11.5	11.7647	3
	PART	1.0	7.7	8.8235	4
	DT	1.0	5.8	7.8431	

Again, if we compare Table 3 and 4, we can notice the feature 109_at of Leukemia dataset has been selected by all three reduction algorithms except boundary-region based (HCBFRBAR) algorithm and 39318_at feature by HCBFREBAR and HCBFRVQLAR algorithm. These two features may be considered as the most significant gene of the dataset.

The fuzzy boundary region-based method finds smaller or equal sized reducts than the other three methods and this is possible because HCBFRBAR includes fuzzy upper approximation information in addition to that of the fuzzy lower approximation.

Table 2 and 3 shows the number of rules generated and overall percentage of classification error obtained using 10-fold cross validation. The classification was performed on the unreduced dataset, followed by the reduced datasets that were obtained using attribute reduction techniques and shown in Table 6 and 7. The performance of all the three classifiers is almost similar for multi-class Leukemia dataset and binary prostate cancer datasets. The minimum and maximum overall classification errors varies from 21.05 ~ 24.56%. In the case of Prostate cancer, minimum and maximum overall classification error varies from 12.75 ~ 17.65%.

Table 6 and 7 shows the percentage of classification error using 10-fold cross validation for ALL, AML and MLL of Leukemia dataset and the percentage of error of tumor and normal sample for Prostate datasets after attribute reduction. Comparing the experimental results of all the three classifiers for both the datasets, it can be seen that the classification results obtained for the reducts generated by hill-climber based fuzzy-rough boundary region attribute reduction (HCBFRBAR) algorithm has outperformed all the remaining results of the classification. In Leukemia dataset, AML has achieved 100% classification accuracy for the combination of HCBFRBAR- JRIP. The least and highest overall classification error ranges from 5.2632 ~ 10.5263%. In Prostate cancer dataset, PART and DT have achieved 99% accuracy for the same reduction algorithm..

The classification accuracy improves significantly for the reduct sets obtained from HCBFRBAR algorithm for both the datasets. JRIP classifier performs better in comparison to the remaining two. This can be attributed to the fact that the proposed method produces smaller subsets for data reduction.

A. Validations of the results

The validation of the findings is shown in two ways i.e., mathematical and real life functional classification of genes. The mathematical validation is based on the induced rules generated from the reducts obtained from different variants of Fuzzy-Rough-Boundary region based reduction algorithms and classification of whole dataset on the basis of that generated rules which have only few responsible genes. The accuracy of prediction of the diseases is verified by applying rule sets generated by JRIP, PART and DT classifier on the datasets. The cross validated result, as shown in Table 8 and 9 indicate that these rules can accurately predict the data. In all the datasets only few marker genes classify the entire datasets which validate our results.

To find biological relevance of the method, next part of validation need to be applied to find actual functional classification of those genes in human body. This is obtained from a Gene Ontology website called DAVID [http://david.abcc.ncifcrf.gov/] where it is available. If the lists of marker genes are provided as input with appropriate gene identifier, the website gives the function of these genes or proteins in human body. In addition to this, it is also possible to find genetic disease which happens due to variation in gene expressions. Table 7 and 8 show functional classification of marker genes for multi-class Leukemia dataset and Prostate dataset.

Table8. Functional Annotations of selected Genes of Leukemia Dataset

Gene-ID	Gene Name and Function obtained from DAVID
35985_at	A kinase (PRKA) anchor protein2; PALM2-AKAP2 Functional annotation: regulation of cell shape, regulation of cell morphogenesis
38017_at	CD79a molecule, immunoglobulin associated alpha Functional annotation: cell activation, immune system development, leukocyte differentiation, positive regulation of immune system process, immune response cell, lymphocyte differentiation, cell proliferation, hemopoiesis, leukocyte activation, lymphocyte activation, leukocyte proliferation
35260_at	MLX interacting protein Functional annotation: chain: MLX-interacting protein, DNA-binding region: Basic motif, domain: Helix-loop-helix motif, domain:Leucine-zipper, modified residue, region of interest: Mediates heterotypic interactions between MLXIP and MLX and is required for cytoplasmic localization, region of interest: Required for cytoplasmic localization, region of interest: Transactivation domain, sequence variant, splice variant
109_at	Rab9 effector protein with kelch motifs Functional annotation: chain:Rab9 effector protein with kelch motifs, repeat:Kelch 1, repeat:Kelch 2, repeat:Kelch 3, repeat:Kelch 4, repeat:Kelch 5, sequence conflict, sequence variant, splice variant,
39649_at	Rho GTPase activating protein 4 Functional annotation: apoptosis, induction of apoptosis, intracellular signalling cascade, Ras protein signal transduction, Rho protein signal transduction, cell death, induction of apoptosis by extracellular signals, regulation of cell death, programmed cell death, death, regulation of apoptosis
39318_at	T-cell leukemia/lymphoma 1A Functional annotation: chain: T-cell leukemia/lymphoma protein 1A, helix, mutagenesis site
650_s_at	Calcium/calmodulin-dependent protein kinase II gamma Functional annotation: active site: proton acceptor, binding site:ATP, chain: calcium/calmodulin-dependent protein kinase type II subunit
34833_at	Family with sequence similarity 32, memberA Functional annotation: Protein of unknown function DUF1754, eukaryotic
306_s_at	High-mobility group neuclosome binding domain 1 Functional annotation: High mobility group protein HMG14 and HMG17
102_at	Homeodomain interacting protein kinase 3 Functional annotation: apoptosis, anti-apoptosis, cell death, regulation of cell death, regulation of stress-activated protein kinase signaling pathway, regulation of cellular response to stress
39155_at	Proteasome (prosome, macropain) 265 subunit, non-ATPase 3 Functional annotation: positive regulation of macromolecule metabolic process, negative regulation of macromolecule metabolic process, modification-dependent protein catabolic process, cell cycle process, protein catabolic process,anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process, regulation of protein ubiquitination,negative regulation of protein ubiquitination, positive regulation of protein ubiquitination, regulation of protein modification process, negative regulation of protein modification process, positive regulation of protein modification process
32715_at	Vesicle associated membrane protein 8 (endobrevin) Functional annotation: protein complex assembly, post-Golgi vesicle-mediated transport, membrane fusion, membrane organization, vesicle-mediated transport, macromolecular complex subunit organization, intracellular transport, Golgi vesicle transport, macromolecular complex assembly, protein complex biogenesis

Table9. Gene functional classification result for Prostrate Dataset

Gene ID	Official Gene Symbol	Gene Function obtained from DAVID
38408_at	TSPAN7	**Disease**: Defects in TSPAN7 are the cause of mental retardation X-linked type 58 (MRX58) [MIM:300210]. Mental retardation is characterized by significantly sub-average general intellectual functioning associated with impairments in adaptative behavior and manifested during the developmental period. Non-syndromic mental retardation patients do not manifest other clinical signs., **Function:** May be involved in cell proliferation and cell motility., similarity:Belongs to the tetraspanin (TM4SF) family., tissue specificity: Not solely expressed in T-cells. Expressed in acute myelocytic leukemia cells of some patients.,
37716_at	CD200	**Function**: Costimulates T-cell proliferation. May regulate myeloid cell activity in a variety of tissues., similarity:Contains 1 Ig-like C2-type (immunoglobulin-like) domain., similarity:Contains 1 Ig-like V-type (immunoglobulin-like) domain., subunit:Interacts with CD200R1.,
37639_at	HPN	**Function**: Plays an essential role in cell growth and maintenance of cell morphology. similarity:Belongs to the peptidase S1 family., similarity:Contains 1 peptidase S1 domain., similarity:Contains 1 SRCR domain., tissue specificity:Present in most tissues, with the highest level in liver.,
38378_at	Cd53	**Function**: May be involved in growth regulation in hematopoietic cells., similarity: Belongs to the tetraspanin (TM4SF) family., tissue specificity:B-cells, monocytes, macrophages, neutrophils, single (CD4 or CD8) positive thymocytes and peripheral T-cells.,

VI. Conclusion

This paper has presented the advantage of the proposed integrated attribute reduction algorithm (HCBFRBAR).The development of fuzzy-rough attribute reduction algorithm is based on the information in the fuzzy boundary region and a heuristic search hill climber to guide the attribute reduction process. When this is minimized, a fuzzy-rough reduct has been achieved. Another three fuzzy-rough hill-climber based attribute reduction algorithms have been used for attribute reduction and the efficiency is compared with the boundary-based hill-climber attribute method. Further research in this area will include a more in-depth experimental investigation of the proposed method and the impact of the choices of relations, connectives and search methods. The performance may also be improved using evolutionary algorithms as search method. This could be achieved by considering the properties of the fuzzy connectives and removing clauses that are redundant in the presence of others.

VII. References

[1]. J. Derisi, L. Penland, P.O. Brown. Use of a cDNA Microarray to Analyse Gene Expression Patterns in Human Cancer, Nature Genetics 14 , 457-460, 1996.

[2]. H. Liu and H. Motoda. Feature Selection for Knowledge Discovery and Data Mining. Boston: Kluwer Academic, 1998.

[3]. I.Guyon and A. Elisseeff. An introduction to variable and feature selection. Journal of Machine Learning Research, 3:1157-1182, 2003.

[4]. H. Liu and H. Motoda, editors. Computational Methods of Feature Selection. Chapman and Hall/CRC Press, 2007.

[5]. Zadeh LA. Fuzzy sets, Inf Control 8(3):, 338-353, 1965.

[6]. Z. Pawlak. Rough sets, International Journal of Information and Computer Science 11 341_356 , (1982).

[7]. F.F. Xu, D.Q. Miao and L. Wei. Fuzzy-Rough Attribute Reduction via Mutual Information with an Application to Cancer Classification, Computers and Mathematics with Applications 57, 1010-1017, 2009.

[8]. R.Jasen and Q. Shen. “Fuzzy-Rough Sets Assisted Attribute Selection”, IEEE Trans. Fuzzy Syst., vol. 15, no. 1, pp. 73-89, 2007.

[9]. Zahra Shaeiri, Reza Ghaderi and Ali Hojjatoleslami,. Fuzzy-Rough Feature Selection and a Fuzzy 2-Level Complementary Approach for Classification of Gene Expression Data, Scientific Research and Essays, vol. 7(14), pp.1512-1500, 16 April, 2012.

[10].Skowron, A., Rauszer, C. The Discernibility Matrices and Functions in Information Systems. In: Slowinski, R. (Eds): Intelligent Decision Support- Handbook of Applications and Advances of the Rough Sets Theory, Kluwer Academic Publishers, Dordrecht, pp. 311-362, 1992.

[11].Hu, X. Knowledge Discovery in Databases: An Attribute Oriented Rough Set Approach, PhD thesis, Regina university, 1995.

[12].Wang, G.Y. Zhao, J. Theoretical Study on Attribute Reduction of Rough Set Theory: Comparison of Algebra and Information Views. Proceedings of the Third IEEE International Conference on Cognitive Informatics, 2004.

[13].Hu, K, Lu, Y. C. Shi, C. Y. Feature Ranking in Rough Sets, AI communication, 16(1), 41-50, 2003.

[14].Zhai, L. Y., et al. Feature Extraction Using Rough Set Theory and Genetic Algorithms- An Application for the Simplification of Product Quality Evaluation. Computers & Industrial Engineering, 43, 661-676, 2002.

[15].Q. Shen and A. Chouchoulas. A fuzzy-Rough Approach For Generating Classification Rules, Pattern Recognit., vol. 35, no. 11, pp. 341-354, 2002.

[16].Jensen. R, Shen, Q. New Approaches to Fuzzy-Rough Feature Selection, IEEE Transaction on Fuzzy Systems, vol., 17, No. 4.pp. 824-838, 2009.

[17].Scott A. Armstrong, et al. "MLL Translocations Specify A Distinct Gene Expression Profile that Distinguishes A Unique Leukemia" Nature Genetics, 30, pp.41-47, 2002.

[18].Dinesh Singh, et al. Gene Expression Correlates of Clinical Prostate Cancer Behavior . Cancer Cell, 1:203-209, March, 2002.

[19].R. Jensen and Q. Shen. Fuzzy-Rough Sets Assisted Attribute Selection, IEEE Trans. Fuzzy syst., vol.15, no. 1 pp. 73-89, 2007.

[20].I. H. Witten and E. Frank. Generating Accurate Rule Sets Without Global Optimization, in Proc. 15th Int. Conf. Mach. Learn. San Francisco, CA: Morgan Kaufmann, 1998.

[21].Ron Kohavi. The Power of Decision Tables. In: 8th European Conference on Machine Learning, 174-189, 1995.

[22].Z. Pawlak. Rough Sets, Int. J. Inform. Comput. Sci. 11, 341-356,1982.

[23].J. Dai, W. Wang, H. Tian, L.Liu. Attribute selection based on a new conditional entropy for incomplete decision systems, Knowledge Based Systems, vol. 39, pg. 207-213, Feb 2013.

[24].J.M. Cadenas, M.C. Garrido, R. Martinez. Feature subset Selection Filter-Wrapper based on low quality data, Expert Systems with Application, vol. 40, Issue 16, 15, Pages 6241-6252, November 2013.

Author Biographies

Sujata Dash: She has received her Ph.D. degree in Computational Modeling and Simulation from Berhampur University, Orissa, India in 1995. She is working as an Associate Professor in Computer Application at North Orissa University, Baripada, Odisha, India and has published more than 57 technical papers in national/ international journals / Proceedings of international conferences / book chapters of reputed publications. Her current research interests include Machine Learning, Data Mining, Bioinformatics, Intelligent Agent, Web Data Mining, Image Processing and Cloud Computing.