
A COMPARATIVE STUDY ON FEATURE REDUCTION ALGORITHMS

JOGA GANGADHAR

Abstract

One of the many successful applications of rough set theory has been to this area. The rough set ideology of using only the supplied data and no other information has many benefits in feature selection, where most other methods require supplementary knowledge. However, the main limitation of rough set-based feature selection in the literature is the restrictive requirement that all data is discrete and it returns first minimal dataset. So, we introduced Decision Relative Discernibility Matrix Approach for finding all possible reduct sets. In classical rough set theory, it is not possible to consider real-valued or noisy data. This thesis proposes and develops an approach based on fuzzy-rough sets, fuzzy rough feature selection (FRFS), that addresses these problems and retains dataset semantics. In the experimental studies, FRFS is shown to equal or improve classification accuracy when compared to the results from unreduced data. Classifiers that use a lower dimensional set of attributes which are retained by fuzzy-rough reduction outperform those that employ more attributes returned by the existing crisp rough reduction method. In addition, it is shown that FRFS is more powerful than the other FS techniques in the comparative study.

Keywords:

Data sets;
ideology;
Decision;
FRFS;
Classification;
Quick Reduct;
Rough Set.

Author correspondence:

JOGA GANGADHAR,
MTECH – COMPUTER SCIENCE & TECHNOLOGY, BITS VIZAG
Udayana University, Jalan P.B.Sudirman, Denpasar, Bali-Indonesia

1. Introduction

Feature Reduction Algorithms

There are many factors that motivate the inclusion of a Feature Reduction step in a variety of problem - solving systems. Many application problems process data in the form of a collection of real-valued vectors (for example, text classification, bookmark categorization, data mining, machine learning, pattern recognition and signal processing). If these vectors exhibit a high dimensionality, then processing becomes infeasible. Therefore, it is often useful, and sometimes necessary, to reduce the data dimensionality to a more manageable size with as little information loss as possible. Sometimes, high-dimensional complex phenomena can be governed by significantly fewer, simple variables. The process of dimensionality reduction here will act as a tool for modelling these phenomena, improving their clarity. There is often a significant amount of redundant or misleading information present; this will need to be removed before any further processing can be carried out. Feature selection is the process of choosing a subset of features from the original set of features forming patterns in a given dataset. The subset should be necessary and sufficient to describe target concepts, retaining a suitably high accuracy in representing the original features. The importance of feature selection is to reduce the problem size and resulting search space for learning algorithms. It can also improve the quality and speed of classification. Due to the abundance of noisy, irrelevant or misleading features, the ability to handle imprecise and inconsistent information in real world problems has become one of the most important requirements for feature selection. Feature Reduction algorithms attempt to reduce the number of dimensions considered in a task so as to improve performance on some dependent measures. [1]

Rough Sets

The problem of imperfect knowledge has been tackled for a long time by philosophers, logicians and mathematicians. Recently it became also a crucial issue for computer scientists, particularly in the area of artificial intelligence. There are many approaches to the problem of how to understand and manipulate the imperfect knowledge. Rough sets can handle uncertainty and vagueness, discovering patterns in inconsistent data. The rough set approach to feature selection is to select a subset of features (or attributes), which can predict the decision concepts as well as the original feature set. The optimal criterion for rough set feature selection is to find shortest or minimal reducts while obtaining high quality classifiers based on the selected features there are many rough set algorithms for feature selection. The most basic solution to finding minimal reducts is to generate all possible reducts and choose any with minimal cardinality, which can be done by constructing a kind of discernibility function from the dataset and simplifying it. Obviously, this is

an expensive solution to the problem and is only practical for very simple datasets. It has been shown that finding minimal reducts or all reducts are both NP-hard problems. Therefore, heuristic approaches have to be considered.[4,5]

Fuzzy - Rough Sets

However, the main limitation of Rough set-based attribute selection in the literature is the restrictive requirement that all data is discrete. In classical rough set theory, it is not possible to consider real-valued or noisy data. The Rough Set Reduction processed discussed previously can only operate effectively with datasets containing discrete values. Additionally, there is no way of handling noisy data. As most datasets contain real-valued attributes, it is necessary to perform a discretization step beforehand. This is typically implemented by standard fuzzification techniques enabling linguistic labels to be associated with attribute values. It also aids the modelling of uncertainty in data by allowing the possibility of the membership of a value to more than one fuzzy label.[7]

2. LITERATURE SURVEY

“Combining information with extraction genetic algorithms for text mining”, provided the basic need of feature selection in Knowledge Discovery in Databases process, importance of feature selection in various application areas and also provides the taxonomy of dimensionality reduction approaches with mentioning the differences between the filter based approach and wrapper based approach.[1]

“Fuzzy-Rough attribute reduction with application to web categorization”, This paper has been concerned with the development of fuzzy-rough attribute selection, which alleviates important problems encountered by tradition Rough set attribute reduction such as dealing with noise and real valued attributes.[6]

“New Approaches to Fuzzy-Rough Feature Selection”, this paper has presented three new techniques for FRFS based on the use of fuzzy T-transitive similarity relations that alleviate problems encountered with FRFS. [7]

“Rough set theory and its applications”, In this paper the basic concepts of rough set theory and its applications to drawing conclusions from data are discussed and for the sake of illustration, an example of churn modeling in telecommunications is presented.[4,5]

“A Rough Set approach to feature selection based on ant colony optimization”, this paper discusses the short comings of conventional hill climbing rough set approaches to feature selection which fails to find optimal reducts, as no perfect heuristic can guarantee optimality. So, Ant Colony Optimization (ACO) approach is provided as a promising feature selection mechanism.

Explaining research model, theory, technique of collecting the data, technique of analyzing the data, hypothesis.research chronological, including research design, research procedure (in the form of algorithms, Pseudocode or other), how to test and data acquisition [1]-[3]. The description of the course of research should be supported references, so the explanation can be accepted scientifically [2], [4].

3. Results and Analysis

INPUT DATASET:

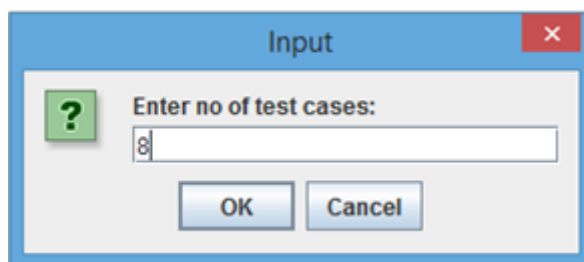
```
# 1  CONDITIONAL  ATTRIBUTE  A
# 2  CONDITIONAL  ATTRIBUTE  B
# 3  CONDITIONAL  ATTRIBUTE  C
# 4  CONDITIONAL  ATTRIBUTE  D
# 5  CONDITIONAL  ATTRIBUTE  E
# 6  DECISION     ATTRIBUTE  F

2, 3, 2, 1, 0, 2
3, 3, 2, 1, 0, 2
2, 2, 2, 2, 1, 2
0, 2, 2, 1, 2, 0
0, 1, 1, 1, 0, 0
1, 1, 1, 1, 1, 1
3, 3, 3, 2, 0, 3
3, 3, 3, 0, 0, 3
```

1SAMPLE DATASET-1

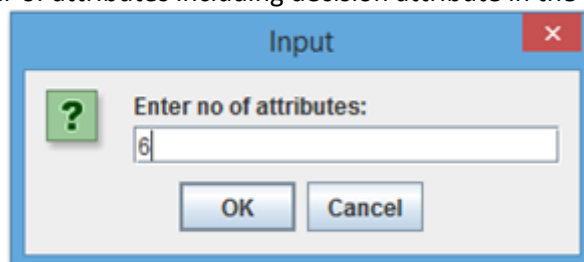
OUTPUT RESULTS:

1. Enter the total number of Test cases i.e (objects) taken in the dataset .



2NO. OF TEST CASES FOR QUICK REDUCT AND MATRIX APPROACH

2. Enter the total number of attributes including decision attribute in the dataset.



3NO. OF TEST CASES FOR QUICK REDUCT AND MATRIX APPROACH

3. Now click on **upload table and Calculate** for finding the minimal data set

A	B	C	D	E	F
2	3	2	1	0	2
3	3	2	1	0	2
2	2	2	2	1	2
0	2	2	1	2	0
0	1	1	1	0	0
1	1	1	1	1	1
3	3	3	2	0	3
3	3	3	0	0	3

4 INPUT DATASET FOR QUICK REDUCT AND MATRIX APPROACH

4. After clicking on that button, the output window shows the all possible reduct sets based on the Boolean discernibility function for the input dataset.

```

Output - MajorProject (run) #4 x
run:
{A∨B∨C}∧{A∨E}∧{C∨D}∧{A∨C∨D}∧{A∨B∨E}∧{A∨B∨C∨D}∧{A∨D∨E}∧{A∨B∨C∨E}∧{A∨B∨C∨D∨E}
After Removing Super Sets:
{A∨B∨C}∧{A∨E}∧{C∨D}
    
```

5 RESULT OF DISCERNIBILITY MATRIX APPROACH

5. After clicking on that button, the output window shows the all possible reduct sets based on the Quick Reduct algorithm for the input dataset.

```

Output - MajorProject (run) x
run:
Reduct set: {
A, C, }
    
```

FUZZY ROUGH QUICK REDUCT RESULTS

INPUT DATASET

A	B	C	D	E	F
2	3	2	1	0	2
3	3	2	1	0	2
2	2	2	2	1	2
0	2	2	1	2	0
0	1	1	1	0	0
1	1	1	1	1	1
3	3	3	2	0	3
3	3	3	0	0	3

6 INPUT DATASET FOR FUZZIFICATION

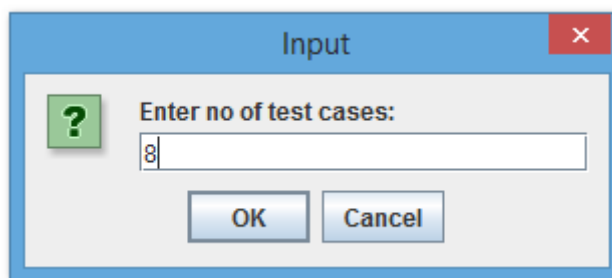
OUTPUT RESULTS :

1. For the above input dataset the corresponding membership values are

Output - project3 (run) x		Tasks													
run:															
	0.2	0.8	0.0	0.6	0.2	0.8	0.8	0.2	0.6	0.0	0.2	0.8			
	0.0	0.6	0.0	0.6	0.2	0.8	0.8	0.2	0.6	0.0	0.2	0.8			
	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.8	0.2	0.2	0.8			
	0.6	0.0	0.2	0.8	0.2	0.8	0.8	0.2	0.2	0.8	0.6	0.0			
	0.6	0.0	0.8	0.2	0.8	0.2	0.8	0.2	0.6	0.0	0.6	0.0			
	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2			
	0.0	0.6	0.0	0.6	0.0	0.6	0.2	0.8	0.6	0.0	0.0	0.6			
	0.0	0.6	0.0	0.6	0.0	0.6	0.6	0.0	0.6	0.0	0.0	0.6			
Reduct Set:															

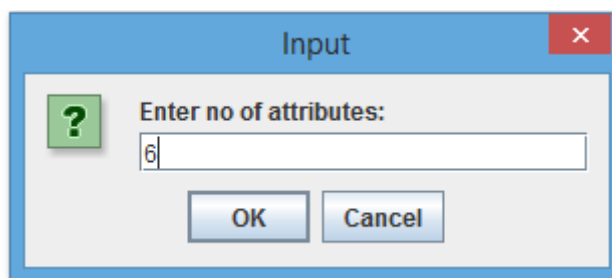
7MEMBERSHIP VALUES OF INPUT DATASET

2. After getting the membership values. Enter the total number of Test cases i.e (objects) taken in the dataset.

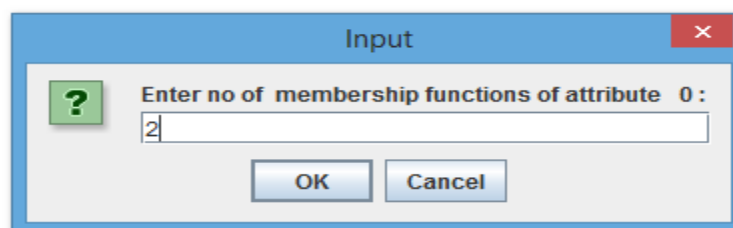


8NO. OF TEST CASES FOR FRFS

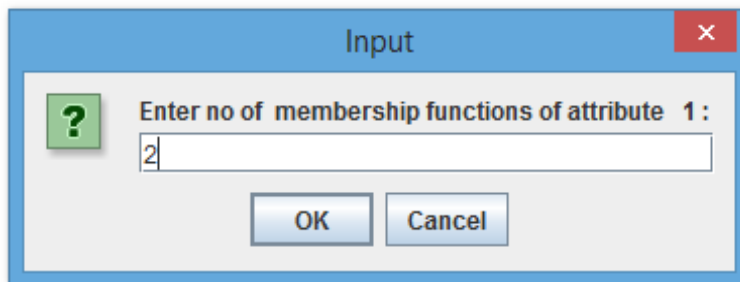
3. Enter the total number of attributes including decision attribute in the dataset.



4. Enter the number of membership functions for every attribute.

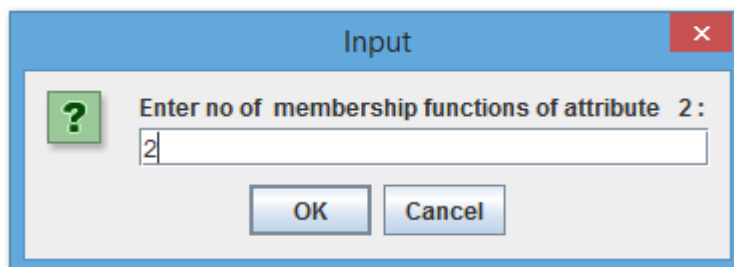


9NO. OF MEMBERSHIP FUNCTIONS FOR A



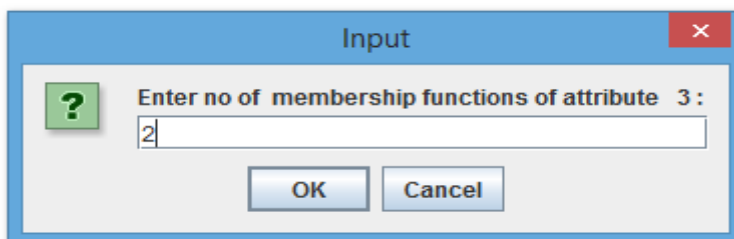
The dialog box has a blue title bar with the text "Input" and a red close button. The main area has a light gray background. On the left, there is a green square with a white question mark. To its right, the text "Enter no of membership functions of attribute 1 :" is displayed. Below this text is a white text input field containing the number "2". At the bottom of the dialog, there are two buttons: "OK" and "Cancel".

10NO. OF MEMBERSHIP FUNCTIONS FOR B



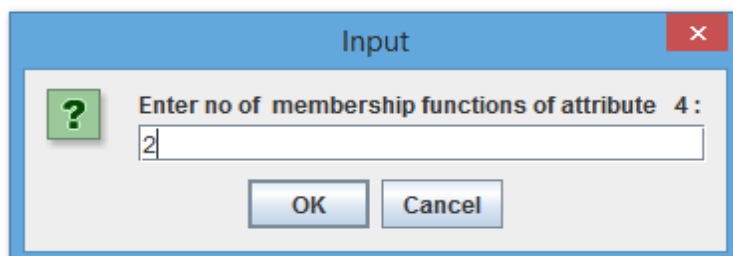
The dialog box has a blue title bar with the text "Input" and a red close button. The main area has a light gray background. On the left, there is a green square with a white question mark. To its right, the text "Enter no of membership functions of attribute 2 :" is displayed. Below this text is a white text input field containing the number "2". At the bottom of the dialog, there are two buttons: "OK" and "Cancel".

11NO. OF MEMBERSHIP FUNCTIONS FOR C



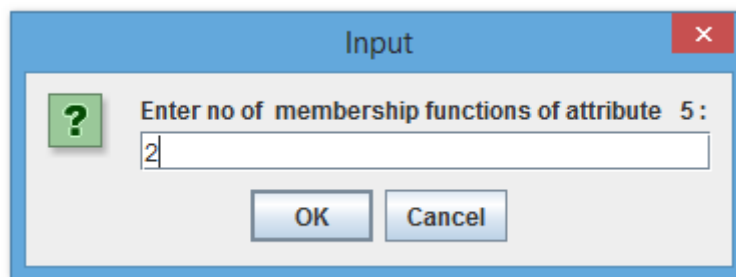
The dialog box has a blue title bar with the text "Input" and a red close button. The main area has a light gray background. On the left, there is a green square with a white question mark. To its right, the text "Enter no of membership functions of attribute 3 :" is displayed. Below this text is a white text input field containing the number "2". At the bottom of the dialog, there are two buttons: "OK" and "Cancel".

12NO. OF MEMBERSHIP FUNCTIONS FOR D



The dialog box has a blue title bar with the text "Input" and a red close button. The main area has a light gray background. On the left, there is a green square with a white question mark. To its right, the text "Enter no of membership functions of attribute 4 :" is displayed. Below this text is a white text input field containing the number "2". At the bottom of the dialog, there are two buttons: "OK" and "Cancel".

13NO. OF MEMBERSHIP FUNCTIONS FOR E



14NO. OF MEMBERSHIP FUNCTIONS FOR F

```

Output - project3 (run) x Tasks
[ ]
[0]
[0] : 0.6
[1]
[1] : 0.30000000000000004
[0]
[0] : 0.6
[2]
[2] : 0.30000000000000004
[0]
[0] : 0.6
[3]
[3] : 0.3
[0]
[0] : 0.6
[4]
[4] : 0.25
[0]
[0] : 0.6
[0, 1]
[0, 1] : 0.6
[0]
[0] : 0.6
[0, 2]
[0, 2] : 0.6
[0]
[0] : 0.6
[0, 3]
[0, 3] : 0.6
[0]
[0] : 0.6
[0, 4]
[0, 4] : 0.6
[0]
[0] : 0.6
[A]
BUILD SUCCESSFUL (total time: 10 seconds)

```

FINAL RESULT - FRFS

WORKING OUT WITH PRACTICAL DATASET

IRIS DATASET

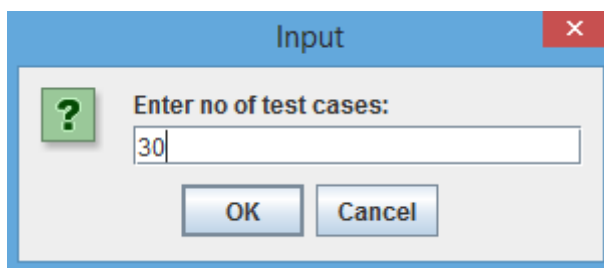
Relevant Information:

This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda & Hart, for example.) The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other. This is an exceedingly simple domain.

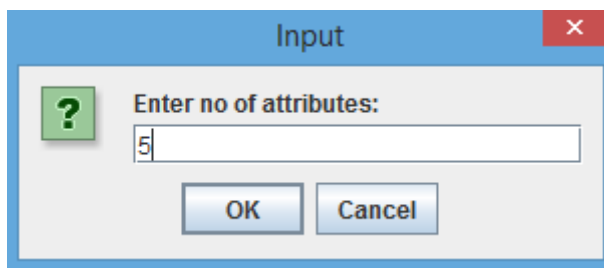
- ✓ Number of Instances: 30 (10 in each of three classes)
- ✓ Predicted attribute: class of iris plant.
- ✓ Number of Attributes: 4 numeric, predictive attributes and the class

Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
 - Iris Setosa
 - Iris Versicolour
 - Iris Virginica



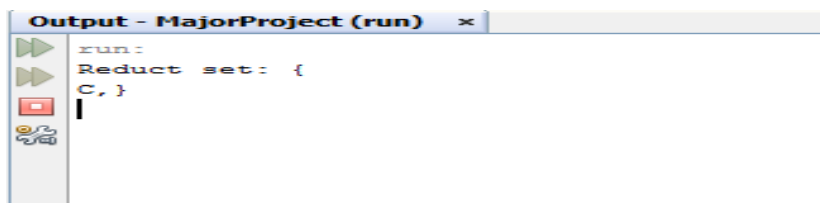
15NO. OF TEST CASES



16NO. OF ATTRIBUTES

A	B	C	D	E
5.4	3.7	1.5	0.2	Irissetosa
4.8	3.4	1.6	0.2	Irissetosa
4.8	3.0	1.4	0.1	Irissetosa
4.3	3.0	1.1	0.1	Irissetosa
5.8	4.0	1.2	0.2	Irissetosa
5.7	4.4	1.5	0.4	Irissetosa
5.4	3.9	1.3	0.4	Irissetosa
5.1	3.5	1.4	0.3	Irissetosa
5.7	3.8	1.7	0.3	Irissetosa
5.1	3.8	1.5	0.3	Irissetosa
5.0	2.0	3.5	1.0	Irisversicolor
5.9	3.0	4.2	1.5	Irisversicolor
6.0	2.2	4.0	1.0	Irisversicolor
6.1	2.9	4.7	1.4	Irisversicolor
5.6	2.9	3.6	1.3	Irisversicolor
6.7	3.1	4.4	1.4	Irisversicolor
5.6	3.0	4.5	1.5	Irisversicolor
5.8	2.7	4.1	1.0	Irisversicolor
6.2	2.2	4.5	1.5	Irisversicolor
5.6	2.5	3.9	1.1	Irisversicolor
6.5	3.2	5.1	2.0	Irisvirginica
6.4	2.7	5.3	1.9	Irisvirginica
6.8	3.0	5.5	2.1	Irisvirginica
5.7	2.5	5.0	2.0	Irisvirginica
5.8	2.8	5.1	2.4	Irisvirginica
6.4	3.2	5.3	2.3	Irisvirginica
6.5	3.0	5.5	1.8	Irisvirginica
7.7	3.8	6.7	2.2	Irisvirginica
7.7	2.6	6.9	2.3	Irisvirginica
6.0	2.2	5.0	1.5	Irisvirginica

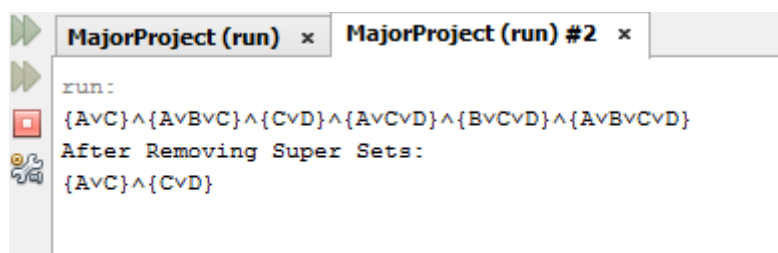
17INPUT DATASET OF IRIS

OUTPUT FOR QUICK REDUCT ALGORITHM


```

Output - MajorProject (run) x
run:
Reduct set: {
C, }

```

*18 RESULT OF RSAR FOR IRIS DATASET***OUTPUT FOR DISCERNIBILITY MATRIX APPROACH**


```

MajorProject (run) x MajorProject (run) #2 x
run:
{AVC}^{AVBVC}^{CVD}^{AVCVD}^{BVCVD}^{AVBVCVD}
After Removing Super Sets:
{AVC}^{CVD}

```

19 RESULT OF MATRIX APPROACH FOR IRIS DATASET

Hence from the above Iris dataset , Quick reduct algorithm results {c} which is minimal subset whereas Discernibility matrix approach results in all possible reduct sets {AC} or {CD} where the intersection results in Core attribute {C}.

4. Conclusion

A summary of the research presented in this dissertation is given, with a focus on the main contribution, Rough set feature selection, fuzzy-rough feature selection, and its applications. Based on a critical survey of the existing literature, it was seen to be the case that many feature selection methods rely on a preliminary discretization procedure in an attempt to deal with noisy and real-valued data. This is particularly the case with rough set-based approaches which depend entirely on crisp datasets. Although this provides a makeshift solution to the problem, it is reliant upon a good discretization that incorporates noise-elimination to produce a useful resulting data reduction. In fact, there may be situations where a dataset contains both nominal and real-valued conditional feature

5 .References

- [1] J. Atkinson-Abutridy, C. Mellish and S. Aitken. Combining information with extraction genetic algorithms for text mining. *IEEE Intelligent Systems*, Vol. 19, No. 3, pp. 22–30. 2004.
- [2] I. D'untsch and G. Gediga. *Rough Set Data Analysis: A road to non-invasive knowledge discovery*. Bangor: Methodos. 2000.
- [3] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth. From data mining to knowledge discovery in databases. *Artificial Intelligence Magazine*, Vol. 17, No. 3, pp. 37–54. 1996.
- [4] Z. Pawlak. *Rough Sets: Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publishing, Dordrecht. 1991.
- [5] Z. Pawlak. Rough Sets. *International Journal of Computer and Information Sciences*, Vol. 11, No. 5, pp. 341–356. 1982.
- [6] Richard Jensen, Qiang Shen. Fuzzy rough attribute reduction with application to web categorization, *Fuzzy Sets and Systems* 141 (2004) 469– 485
- [7] Richard Jensen, Qiang Shen. New Approaches to Fuzzy-Rough Feature Selection, *IEEE Transactions On Fuzzy Systems*, Vol. 17, No. 4, August 2009.
- [8] Kantardzic Mehmed (2001) *Data Mining Concepts, Models, Methods, and Algorithms*.
- [9] Acuña, E. (2003). A comparison of filters and wrappers for feature selection in supervised classification. *Proceedings of the Interface 2003 Computing Science and Statistics*.
- [10] Guyon, I, and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*.
- [11] Kohavi, R. (1994). A third dimension to rough sets. In *proceeding of the Third international workshop on rough sets and soft computing*.
- [12] Ling Partee *Basic Concepts of Set Theory, Functions and Relations*. adapted from lecture notes 409,.
- [13] *Intelligent Systems and controls Module -2 Lecture -1 Fuzzy Sets : A primer* by Prof. Laxmidhar Behera. (NPTEL Lecture video).
- [14] W. Siedlecki and J. Sklansky. A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters*, Vol. 10, No. 5, pp. 335–347. 1989.
- [15] R. Jensen and Q. Shen. Rough and fuzzy sets for dimensionality reduction. In *Proceedings of the 2001 UK Workshop on Computational Intelligence*, pp. 69–74. 2001.

- [16] R. Jensen and Q. Shen. Fuzzy-Rough Sets for Descriptive Dimensionality Reduction. In Proceedings of the 11th International Conference on Fuzzy Systems, pp. 29–34. 2002.
- [17] M. Dash and H. Liu. Feature Selection for Classification. *Intelligent Data Analysis*, Vol. 1, No. 3, pp. 131–156. 1997.
- [18] H. Sever. The status of research on rough sets for knowledge discovery in databases. In Proceedings of the Second International Conference on Nonlinear Problems in Aviation and Aerospace (ICNPAA98), Vol. 2, pp. 673–680. 1998.
- [19] J. Jelonek and J. Stefanowski. Feature subset selection for classification of histological images. *Artificial Intelligence in Medicine*, Vol. 9, No. 3, pp. 227–239. 1997.