International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 8, ISSN No. 2455-2143, Pages 303-306
Published Online June - July 2016 in IJEAST (http://www.ijeast.com)

APPLICATION OF FUZZY-ROUGH SET THEORY FOR FEATURE SUBSET SELECTION

Surekha Samsani
Department of CSE
JNTUK-UCEV, Vizianagaram,
AP, India

Dr. G. Jaya Suma
Department of IT
JNTUK-UCEV, Vizianagaram,
AP, India

***Abstract*— Fuzzy Set Theory and Rough Set Theory are the most popular mathematical tools for dealing with uncertainties. During past decades, these set theories are being applied successfully in several areas for solving many complex tasks. This paper is concerned with the application of hybrid Fuzzy-Rough set based approach for feature subset selection.**

Keywords— **Fuzzy set theory, Rough Set theory, Fuzzy-Rough set theory, Feature Subset Selection.**

I. INTRODUCTION

In real world, knowledge is represented as an information system with set of objects and each object is represented by a set of features also called as attributes. Not all features are necessary for characterizing an object for a particular task. In such circumstances, using all the features increases both the time and space complexity. Hence, it is required to select only relevant set of features that are required for a particular task without affecting the efficiency of the required task. During several decades, Rough Set Theory(RST) is being applied successfully for selecting the most promising set of features especially, when the data is consisting of inconsistencies and uncertainties. RST[1] is a powerful intelligent mathematical tool developed by Pawlak to deal with inconsistencies but, RST can perform well only when the data is discrete. Many real world datasets consist of continuous values. So, in order to apply RST to deal with uncertainty, data should be discretized beforehand. In this paper, standard fuzzification techniques of Fuzzy set theory[2] have been applied to transform continuous values to discrete ones because, fuzzification also aids in dealing with uncertainties by allowing the possibility of the membership degrees to more than one fuzzy label. Hence, this Fuzzy-Rough set based approach for feature selection can perform better than pure RST based feature selection techniques.

The rest of the paper is structured as follows: section II discusses the methodology of the work, and section III gives experimental results and finally section IV concludes the paper.

II. METHODOLOGY

The methodology of the proposed work is shown in figure 1.

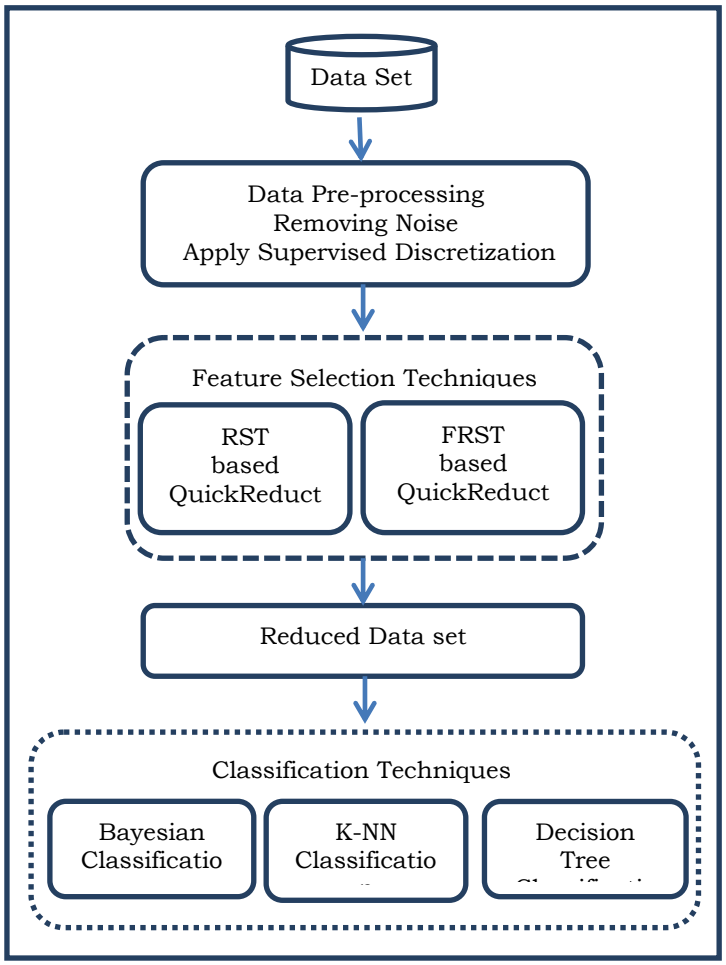

Fig. 1 Methodology

The methodology of the proposed work can be divided into three major phases namely; Preprocessing, RST based and FRST feature selection and then evaluating the performance of the obtained feature subsets by submitting to the classification techniques.

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 8, ISSN No. 2455-2143, Pages 303-306
Published Online June - July 2016 in IJEAST (http://www.ijeast.com)

Phase 1: Preprocessing of the Input Dataset

Real world data sets are very susceptible to noise and consist of many missing values and inconsistencies. So it is required to preprocess the data by filling in missing values and removing inconsistencies. Inconsistent data affects the accuracy of many data mining algorithms. In this paper, inconsistent data objects were removed using RST concepts and then applied supervised discretization [3] technique to discretize all continuous values in the given dataset.

Phase 2: RST and FRST based Feature Selection

The basic concepts of Rough Set theory and Fuzzy-Rough Set theory can be found in [4-9]. The basic Quick Reduct algorithm [10] for reduct generation is given below.

Quick Reduct Algorithm

Input: CA, the set of Conditional attributes
DA, the set of Decision attributes
Output: R, Reduct with minimal set of Conditional attributes

(1) R← {}, empty set
(2) Do
(3) TR←R
(4) ∀ A ϵ (CA-R)
(5) If $\gamma_{(R \cup A)}(DA) > \gamma_{TR}(DA)$ then
(6) TR←R ∪A
(7) R←TR
(8) Until $\gamma_R(DA) = \gamma_{CA}(DA)$
(9) Return R

Quick Reduct algorithm works by adding one attribute at a time from the given set of conditional attributes only, when there is an improvement in the dependency measure and terminates when the addition of any new attribute doesn't improves the dependency measure.
Fuzzy-Rough feature selection techniques use the Fuzzy-Rough dependency which can be derived from Fuzzy Lower approximation [7].

RST and FRST based feature selection techniques are implemented in R STUDIO[11] and the code is given below.

```
## Discretization
myquickreduct<-
function(directory,da,discretization,quickreduct){
library(RoughSets)

a<-read.csv(directory,na.strings="NA",header=FALSE)

mytable<-SF.asDecisionTable(dataset=a,decision.att=da)
#shuffle the data with set.seed
dt.shuffled<-mytable[sample(nrow(mytable))]
#split the data into training and testing
#80% for training and 20% for testing
idx<-round(0.8* nrow(dt.shuffled))
data.tra<-SF.asDecisionTable(dt.shuffled[1:idx,],
                  decision.attr=da,indx.nominal=da)
data.tst<-SF.asDecisionTable(dt.shuffled[(idx+1):
                  nrow(dt.shuffled),- ncol(dt.shuffled)])

##FRST Based Feature Selection
b.frst<-FS.feature.subset.computation(data.tra,
                  method="quickreduct.frst")
print(b.frst)
fs.tra<-SF.applyDecTable(d.tra,b.rst)
##Write it to table
write.csv(d.tra,"F:/discretization.csv")
write.csv(fs.tra,"F:/quickreduct.csv")

##RST Based Feature Selection
b.rst<-FS.feature.subset.computation(d.tra,
                  method="quickreduct.rst")
print(b.rst)
fs.tra<-SF.applyDecTable(d.tra,b.rst)
write.csv(d.tra,"discretization")
write.csv(fs.tra,"quickreduct")
obj.MV<-MV.missingValueCompletion(mytable,
                  type.method="deletionCases")
## Discretization
cut.values<-D.global.discernibility.heuristic.RST(data.tra)
d.tra<-SF.applyDecTable(data.tra,cut.values)
d.tst<-SF.applyDecTable(data.tst,cut.values)
##Feature Selection
b.frst<-FS.feature.subset.computation(data.tra,
                  method="quickreduct.frst")
print(b.frst)
fs.tra<-SF.applyDecTable(d.tra,b.rst)
```

Phase 3: Evaluating the Performance of the obtained Feature Subsets

To evaluate the performance of the Feature subsets generated by the RST based and FRST based feature selection techniques were submitted to three classification techniques namely K-Nearest Neighbor[12], Naïve Bayes[13] and Decision Tree[14].

III. EXPERIMENT AND RESULT

Datasets are taken from UCI machine learning repository [15] and the description of the datasets are given below in Table.1.

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 8, ISSN No. 2455-2143, Pages 303-306
Published Online June - July 2016 in IJEAST (http://www.ijeast.com)

Table 1 Description of the Dataset

The observed classification accuracies for these undiscretized datasets are given in Table 2.

Data set	No. of Attributes	No. of Instances	No. of Classes	Class Names
Breast Cancer	32	569	2	{M,B}
Erythematous – Squamous	35	366	6	{1,2,3,4,5,6}
Hepatitis	20	155	2	{1,2}
Lung Cancer	57	32	3	{1,2,3}
Prognostic	35	198	2	{N,R}
SPECT	23	267	2	{0,1}
SPECTF	45	267	2	{0,1}

Table 2. Accuracy of Classification Techniques on Undiscretized Data

Data Set	Decision Tree (J48)	Bayesian Classification	K-Nearest Neighbor
SPECT	78.5%	73.7%	66.2%
SPECTF	79.4%	62.9%	67.4%
Breast Cancer	92.9%	92.73%	88.9%
Hepatitis	80.0%	83.7%	83.7%
Lung Cancer	81.4%	79.7%	55.5%
Prognostic	73.7%	74.7%	69.1%
Erythematous – Squamous	93.2%	97.7%	96.3%

All continuous valued attributes are discretized and the observed accuracies for various classification techniques on discretized datasets are given in Table 3.

Table 3 Accuracy of Classification Techniques on Discretized Data

Data Set	Decision Tree (J48)	Bayesian Classification	K-Nearest Neighbor
SPECT	68.75%	73.75%	66.25%
SPECTF	79.4%	73.78%	77.90%
Breast Cancer	95.95%	95.58%	96.13%
Hepatitis	80.0%	83.75%	83.75%
Lung Cancer	62.85%	70.37%	55.55%
Prognostic	76.26%	76.26%	76.26%
Erythematous – Squamous	93.29%	98.31%	96.36%

Subsets generated after applying RST based Quick Reduct (RST-QR) and Fuzzy Rough Set based Quick Reduct (FRST-QR) on the above mentioned datasets is given in Table 4.

Table 4. Performance of RST and FRST based Feature Subsets

Data Set	Full Set of Features	RST-QR	FRST-QR
SPECT	23	12	07
SPECTF	45	20	13
Breast Cancer	32	11	07
Hepatitis	20	05	09
Lung Cancer	57	24	06
Prognostic	35	09	06
Erythemato - Squamous	35	17	04

The comparison of the feature subsets generated by RST-QR, and FRST-QR is shown in figure 2.

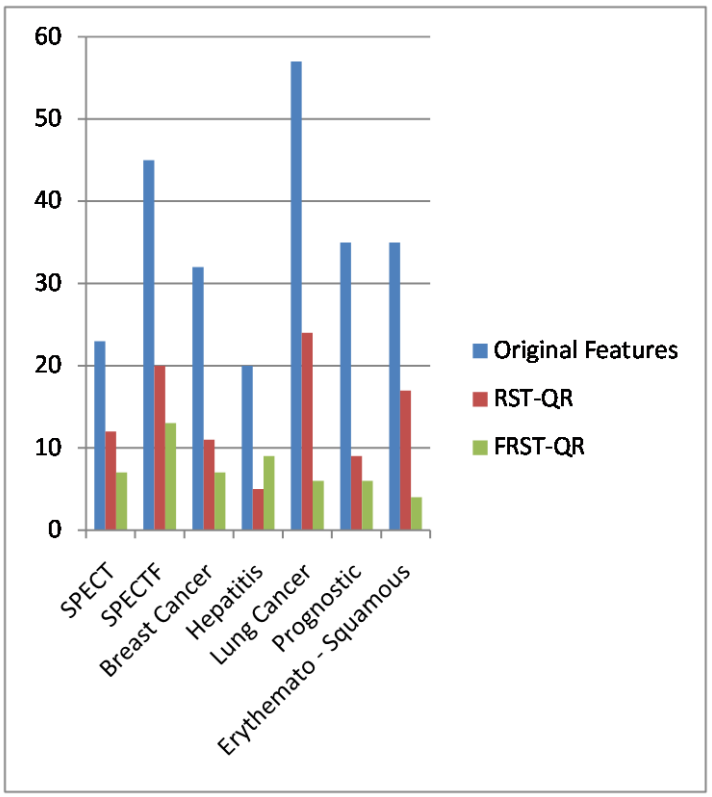

Fig. 2 Performance of Feature Selection Techniques

The observed classification accuracies on the reduced datasets is given in Table 5.

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 8, ISSN No. 2455-2143, Pages 303-306
Published Online June - July 2016 in IJEAST (http://www.ijeast.com)

Table 5. Classification Accuracy of Decision Tree on Reduced Data

Data Set	Full Features	RST-QR	FRST-QR
SPECT	68.75%	70.0%	60.0%
SPECTF	79.4%	78.5%	79.4%
Breast Cancer	95.95%	93.84%	94.9%
Hepatitis	80.0%	90.6%	82.5%
Lung Cancer	62.85%	63.6%	33.3%
Prognostic	76.26%	77.21%	73.4%
Erythemato - Squamous	93.29%	87.7%	66.7%

Table 6. Accuracy of Bayesian Classification on Reduced Data

Data Set	Full Features	RST-QR	FRST-QR
SPECT	73.75%	78.7%	60.0%
SPECTF	73.78%	78.4%	66.8%
Breast Cancer	95.58%	97.58%	94.5%
Hepatitis	83.75%	81.2%	71.2%
Lung Cancer	70.37%	63.6%	48.1%
Prognostic	76.26%	77.21%	74.0%
Erythemato - Squamous	98.31%	91.6%	68.0%

Table 7. Accuracy of K-NN Classification on Reduced Data

Data Set	Full Features	RST-QR	FRST-QR
SPECT	66.25%	70.0%	60.0%
SPECTF	77.90%	79.8%	78.5%
Breast Cancer	96.13%	97.58%	93.4%
Hepatitis	83.75%	87.5%	76.2%
Lung Cancer	55.55%	63.6%	44.4%
Prognostic	76.26%	77.21%	64.5%
Erythemato - Squamous	96.36%	88.1%	68.70%

IV. CONCLUSION

In this paper, RST based Quick Reduct and Fuzzy Rough set theory based Quick Reduct algorithms have been applied on the several bench mark medical datasets for selecting the most promising features. And for testing the efficiency of these techniques, the reduced datasets were submitted to three different classifiers namely Decision Tree, K-Nearest Neighbor and Bayesian classification. Experimental results revealed that, even though the FRST based Quick Reduct generated minimal subset of features; the classification accuracies are not acceptable when compared to the other pure RST based feature selection techniques.

V. REFERENCES

[1] Z.Pawlak, "Rough sets", International Journal of Computer and Information Science11, 341- 356, 1982.

[2] L.A.Zadeh, "Fuzzy sets", Information and control, vol.8,pp.338-353,1965.

[3] H.Liu, F.Hussain,C.L.Tan and M.Dash, "Discretization: An Enabling Technique", Data mining and Knowledge Discovery, Kluwer Academic Publications, Netherlands, 6, pp. 393-423,2002.

[4] Z.Pawlak, "Rough Set approach to Knowledge-based Decision Support", European Journal of Operational Research, 99, pp. 48-57,1997.

[5] Z.Pawlak, "Rough Sets. Theoretical Aspects of Reasoning about Data", Kluwer Academic Publications, Netherlands,1995

[6] Z. Pawlak, "On Rough Dependency of Attributes in Information Systems", Bulletin of the Polish Academy of Sciences, vol.33, pp. 551–599,1985.

[7] R. Jensen, "Combining Rough and Fuzzy sets for Feature Selection", Ph.D. dissertation, School of Informatics, University of Edinburgh, Edinburgh, UK, 2005.

[8] R. Roselin, K. Thangavel, and C. Velayutham, "Fuzzy-Rough Feature Selection for Mammogram Classification" , Journal of Electronic Science and Technology, vol. 9, no. 2, 2011 , 124 -132

[9] D. Dubois and H. Prade, "Putting Rough sets and Fuzzy sets together". In [171], pp. 203–232. 1992.

[10] K.Thangavel and A. Pethalakshmi, "Dimensionality Reduction based on Rough Set Theory", Applied Soft Computing, Elsevier, 9, 1-12.2009.

[11] LalaSeptemRiza , Implementing algorithms of Rough set theory and Fuzzy rough set theory in the R package, ''RoughSets'' , Information Sciences, Elsevier, July,2014,PP: 68-89

[12] P.Cunningham and S.J.Delany "K-Nearest Neighbor Classifiers,‖ Multiple Classifier Systems", Technical Report UCD-CSI-2007, 4,2007.

[13] K.P.Murphy "Naïve Bayes Classifiers", University of British Columbia,2006.

[14] J.R.Quinlan , "C4.5: Programs for Machine Learning", Morgan Kaufmann, 1993.

[15] UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets.html