Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.8 No.S1, 2019, pp. 103-106

An Experimental Analysis on Rough Set Mean, Median, Mode Method of Dependency Values for Feature Selection in Medical Databases

S. Devi[1] and V. Sasirekha[2]
[1]Assistant Professor, Department of MCA, SSM College of Engineering, Tamil Nadu, India
[2]Assistant Professor, Department of CSE, JKK Nataraja College of Arts and Science, Tamil Nadu, India
E-Mail: devikrishnamca@gmail.com, sasirekhailangkumaran@gmail.com

***Abstract* - The problem of imperfect knowledge has been tackled for a long time by philosophers, logicians and mathematicians. Recently it became an important issue for scientists, particularly in the area of Artificial Intelligence. Their square measure several approaches to the matter of the way to perceive and manipulate imperfect information. The most successful approach is based on the rough set notion proposed by Z. Pawlak in the article [1]. The proposed method to find the quick reduct in medical data set using the roughest theory. This method has applied in many classification algorithms and find the measures to calculate the accuracy of this proposed method.**

Keywords: **Rough Set, Dependency Values, Approximation, RST Mean, RST Median, RST Mode**

I. INTRODUCTION

Rough set theory is at a halt approach to vagueness. Similarly, to fuzzy set theory it is not an alternative to conventional set theory but it is embedded in it. Rough set theory can be seen as a specific execution of Frege's idea of vagueness, i.e., imprecision in this approach is expressed by a boundary region of a set, and not by a partial membership, resembling in fuzzy set theory. Rough set model will be outlined quite typically by means that of topological operations, interior and closure, referred to as approximations. Allow us to describe this downside additional exactly. Let X might be a set of U. Characterize the set X with regard to R. to the current finish we are going to need the essential ideas of rough pure mathematics given below.

1. The lower approximation of a set X with respect to R is the set of all objects, which can be classified as X with respect to R
2. The upper approximation of a set X with respect to R is the set of all objects which can be probably classified as X with respect to R.
3. The boundary region of a set X with respect to R is the set of all objects, which can be classified neither as X nor as not-X with respect to R.
4. Set X is crisp, if the boundary region of X is empty.
5. Set X is rough, if the boundary region of X is not empty.

Thus, a set is rough (imprecise) if it has nonempty boundary region; otherwise the set is crisp (precise). This is exactly the ideas of vagueness proposed Frege [9]. The approximations and the boundary region can be defined more precisely. To this end we need some additional notation. The equivalence class of R determined by element x will be denoted by R(x). The indiscernibility relation in certain sense describes our lack of knowledge about the universe. Formal definitions of approximations and the boundary region are as follows:

R-lower approximation of X,

$$\boldsymbol{R}_{*}(\boldsymbol{x}) = \bigcup_{\boldsymbol{x}\in\boldsymbol{U}} \{\boldsymbol{R}(\boldsymbol{x}) : \boldsymbol{R}(\boldsymbol{x}) \subseteq \boldsymbol{X}\}$$

R-upper approximation of X

$$\boldsymbol{R}^{*}(\boldsymbol{x}) = \bigcup_{\boldsymbol{x}\in\boldsymbol{U}} \{\boldsymbol{R}(\boldsymbol{x}) : \boldsymbol{R}(\boldsymbol{x}) \cap \boldsymbol{x} \neq \boldsymbol{\Phi}\}$$

II. LITERATURE REVIEW

There are completely different concepts are applied in feature choice exploitation the roughest theory, [2] applied feature choice in Neighborhood roughest.[3] mention Feature choice exploitation rough set supported criteria.[4] has assumed that one feature sequence is set for all potential object instances, that's next feature within the order doesn't depend upon values of the previous options. The rule is given generating a sequence underneath these conditions. [5] planned a new thought referred to as the "Incremental Dependency Class" (IDC), that calculates the attribute dependency while not exploitation the positive region. [6] Introduced a hypothetical framework supported rough pure mathematics, referred to as positive approximation, which may be accustomed accelerate a heuristic method of attribute reduction.

III. DATA COLLECTION AND PREPROCESSING

Medical data set is taken from UCI repository with 178 features for the quick reduct, this consists of Conditional attribute C(i=1,2,3….n-1) and Decision attribute D(n). Before apply the proposed algorithm for quick reduct the data set should be processed. Here the min max discretization have applied for reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace definite data values. Next to Construct the rough set based equivalence relation for decision attribute separately (i.e) equivalence (i=n).

A. Without Selection Measures

Initially started without any selection of features in the data set directly applied different classification Algorithms Decision Table, JRip, J48, Random Forest, MLP, Naïve Bayes, K-NN and gained the following output[10]. Table I shows the graphical representation of performance comparison with various algorithms also shown in Fig 1.

TABLE I EXPERIMENTAL RESULT OF PERFORMANCE COMPARISON WITHOUT FEATURE SELECTION

Measures	Decision Table	JRip	J48	Random Forest	Naïve Bayes	MLP	K-NN
Precision	0.359	0.448	0.504	0.576	0.432	0.501	0.498
Recall	0.371	0.369	0.503	0.584	0.433	0.502	0.499
F-Measure	0.332	0.278	0.503	0.577	0.416	0.502	0.499

IV. PROPOSED METHOD

This new proposed method consists of three step process. First step is to make the data as perfect for this Min-Max discretization has applied. Second step is to select features the mean, mode, median method is used. Third step is to apply various classification algorithms to find accuracy of this method. The pictorial representation of proposed algorithm shown in Fig 1.

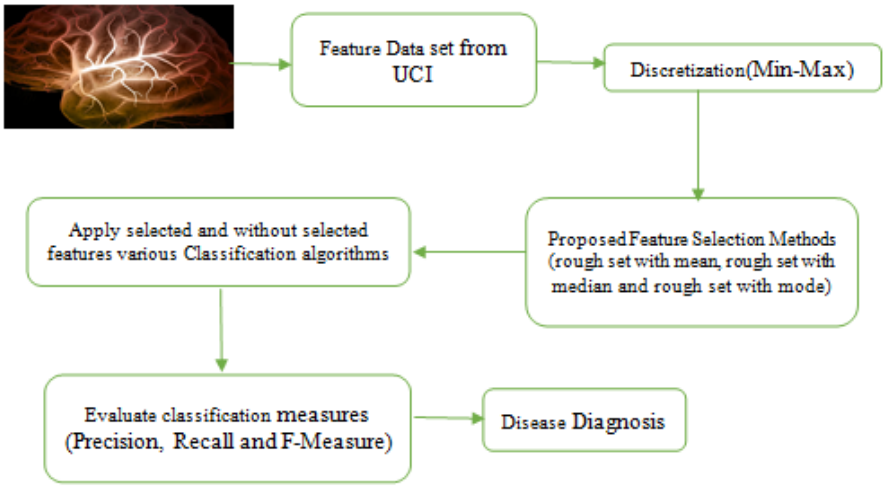

Fig. 1 Pictorial Representation of Proposed Algorithm

V. PROPOSED ALGORITHM

Input: Conditional attribute C(i=1,2,3….n-1) and Decision attribute D(n)
Output: Reduct feature sets (rough set with mean, rough set with median and roughest with mode)
Step 1: Construct the rough set based equivalence relation for each conditional attribute separately (i.e) equivalence (i=1, 2,3….n-1)
Step 2: Construct the rough based equivalence relation for decision attributes separately (i.e) equivalence (i=n).
Step 3: Construct the rough set based dependency value each conditional attribute using decision attribute. (i.e.) equivalence C(i=1,2,3….n-1) divided by D(i=n).
Step 4: Find the geometric mean value from that conditional attribute dependency value to reduct the features using geometric mean.(i.e) conditional dependency >geometric mean
Step 5: Find the geometric median value from that conditional attribute dependency value to reduct the features using geometric median. (i.e.) conditional dependency > geometric median
Step 6: Find the geometric mode value from that conditional attribute dependency value to reduct the features using geometric mode. (i.e) conditional dependency > geometric mode
Step 7: Finally the reduct features are rough set with mean features, roughset with median features and rough set with mode features

A. Calculation of Dependency

For an attribute value system S-{U,CUD,V,F} suppose Ci denotes condition instance in Rc,Dj denotes a decision instance in RD;Rc and RD are the equivalence classes of C and D respectively over U[7]. Then the definition of dependency of decision instance on condition instance is measured by Sij: Ci→ Dj

$$Sij = Card(C_i \cap D_j)/Card(C_i)$$

B. Rough Set with Mode (RSTMode)

The mode of a collection of information prices is that the value that seems most frequently. The point x at that its likelihood mass perform takes its most value. In alternative words, it's the worth that's possibly to be sampled. Like the applied math mean and median, the mode may be a means of expressing, in an exceedingly (usually) single variety, vital info a few variable quantity or a population. The numerical price of the mode is that the same as that of the mean and median in the distribution, and it's going to be terribly completely different in extremely inclined distributions.

TABLE II EXPERIMENTAL RESULT OF PERFORMANCE COMPARISON USING PROPOSED RSTMODE METHOD

Measures	Decision Table	JRip	J48	Random Forest	Naïve Bayes	MLP	K-NN
Precision	0.557	0.64	0.702	0.778	0.646	0.686	0.674
Recall	0.573	0.689	0.705	0.786	0.623	0.692	0.675
F-Measure	0.534	0.60	0.704	0.775	0.618	0.689	0.675

The Table II has displayed the various classification algorithms using RSTMode method of feature selection.

C. Rough Set with Median

The geometric median of a distinct set of sample purposes during a Euclidean space are that the point minimizing the ofdistances to the sample points. This generalizes the median, that has the property of minimizing the of distances for one-dimensional information, and provides a central tendency in higher dimensions.The table 3 has displayed the various classification algorithms using RSTMedian method of feature selection.

TABLE III EXPERIMENTAL RESULT OF PERFORMANCE COMPARISON USING PROPOSED RSTMEDIAN METHOD

Measures	Decision Table	JRip	J48	Random Forest	Naïve Bayes	MLP	K-NN
Precision	0.765	0.819	0.855	0.843	0.813	0.854	0.847
Recall	0.774	0.82	0.866	0.858	0.828	0.858	0.849
F-Measure	0.753	0.818	0.858	0.863	0.822	0.857	0.849

D. Rough Set with Mean

The mean of the dependency values can be calculated and applied the various classification algorithms.Table III displayed the experimental results. Mean is what most people commonly refer to as an average. The mean refers to the number you obtain when you sum up a given set of numbers and then divide this sum by the total number in the set. Mean is also referred to more correctly as arithmetic mean. The mean value can be taken from the dependency values from the dataset. The geometric mean of a data set{a1,a2,……..,an} is given by

$$\left(\prod_{i=1}^{n} a_i\right)^{1/n} = \sqrt[n]{a_1 a_2} \ldots . a_n$$

TABLE IV EXPERIMENTAL RESULT OF PERFORMANCE COMPARISON USING PROPOSED RST MEAN METHOD

Measures	Decision Table	JRip	J48	Random Forest	Naïve Bayes	MLP	K-NN
Precision	0.774	0.821	0.863	0.877	0.837	0.866	0.853
Recall	0.782	0.835	0.885	0.882	0.849	0.874	0.856
F-Measure	0.747	0.824	0.857	0.874	0.839	0.871	0.855

The table IV has displayed the various classification algorithms using RSTMean method of feature selection.

E. Classification Validation and Comparative Analysis

The above proposed algorithm validated by percentage split method. Percentage Split (Fixed or Holdout) is a re-sampling method that leaves out random N% of the original data. For example, select: 75% of the rows formed the training set for building the model. 25% of the rows formed the test set for testing the model. The Precision, Recall and F-Measure taken as measures of accuracy. This proposed method applied many classification algorithms like Decision Table, JRip, J48, RandomForest, MLP, Naïve Bayes, K-NN to find the accuracy. The below tables and graph shown the comparative analysis of various classification algorithms[8]. Finally mean and median has produces 87 features from 178.

TABLE V OUTCOME OF THE PROPOSED ALGORITHMS

Proposed Methods	Total Number of Features	Selected Features using proposed methods
Information gain using rough set with Mean Features	178	87
Information gain using rough set with Median Features	178	87
Information gain using rough set with Mode Features	178	103

VI. CONCLUSION

This paper propose three approaches for rough set primarily based feature choice Mean, Mode, Median supported the dependency values. This approach may be appropriate for the massive information set. From the higher than experiment RSTMean, RSTMedian cause the high accuracy and therefore the RSTMode cause low accuracy. This accuracycan vary depends on the datasets. In future the identical strategies are going to be applied to the various datasets and analyze the accuracy of this technique.

REFERENCES

[1] Z. Pawlak, "Rough set theory and its applications to data analysis", *Cybernetics and Systems*, pp.661-688, 29, Oct. 2010.

[2] C. Wang, M. Shao, Q. He, Y. Qian and Y. Qi, "Feature subset selection based on fuzzy neighborhood rough sets"- *Knowledge-based systems, Elsevier,* Vol. 111, pp. 173-179, Nov. 2016.

[3] N. Zhong, J. Dong and S. Ohsuga "Using rough sets with heuristics for feature selection", *Journal of intelligent information systems, Springer* Vol. 16, No. 3, pp. 199-214, Aug. 2001.

[4] M. Modrzejewski, "Feature selection using rough sets theory", in *Proc. ECML1993, Springer,* pp. 216-226, April 1993.

[5] M. S. Raza and U. Qamar, "An incremental dependency calculation technique for feature selection using rough sets", *Information Sciences, Elsevier,* Vol. 344, pp. 41-65, May 2016.

[6] Y. Qian, J. Liang, W. Pedrycz and C. Dang ," Positive approximation: an accelerator for attribute reduction in rough set theory", *Artificial Intelligence, Elsevier,* Vol. 174, No. 9-10, pp. 597-618, June 2010

[7] Kuo, Tien-Fang, "Approximate Reducts of an Information System," in *Proc. RSFDGrC 2003, Springer*, pp. 291-294, May 2003.

[8] J. Dai and Q. Xu, "Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification", *Applied Soft Computing, Elsevier,* Vol. 13, No. 1, Jan2013.

[9] H. H. Inbarani and S. S. Kumar, "Soft rough sets for heart valve disease diagnosis", in *Proc. AMLTA 2014, Springer,* pp. 347-356, 2014.

[10] Y. Yao and Y. Zhao, "Attribute reduction in decision-theoretic rough set models", *Information sciences, Elsevier,* Vol. 178, No. 17, pp. 3356-3373, Sep. 2008.