International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 2, February - 2014

A Hybrid Technique for Rule generation from Incomplete Quantitative Data.

[1.] Mohsin Mulani, [2.] Mayur More, [3.] Prakash Muchandi , [4.] Ravindra Shewale
[1,2,3,4.] Computer Dept.
Pimpri Cinchwad College of Engineering
Pune, India

***Abstract*-- Data mining is perceived to be the most significant and challenging process. Mining useful information efficiently from large collection of data is definitely important. Generally many proposed algorithms work on complete data sets and it is comparatively easier to perceive knowledge from the same. Rule generation for further analysis of datasets becomes difficult and hence a technique to address completion of datasets becomes a challenging task. In this paper, the proposed algorithm works towards addressing the issue of knowledge extraction from incomplete datasets.**

***Keywords*—Data mining; Quantitative Data; Incomplete Data; Association Rules; Fuzzy Logic; Classification and Rough set theory.**

I. INTRODUCTION

In data mining researches, inducing association rules from datasets is most commonly seen. Most of the previous research works can, however, only handle datasets with attributes of binary values [6]. In real-world applications, datasets are usually composed of quantitative values. Designing data-mining algorithm to deal with different types of data turns a challenge in this research topic. Fuzzy set theory is being used more and more frequently in intelligent systems because of its simplicity and similarity to human reasoning [6]. We are proposing a mining approach that integrated fuzzy logic concepts with Apriori mining algorithm to find interesting item sets and fuzzy association rules in datasets with quantitative values. The term item sets was first proposed by Agrawal et al. in their papers on data mining, and from then becomes a common usage in this field [2]. It means a set composed of items. This paper proposes another new fuzzy mining algorithm based on the Apriori approach to find fuzzy association rules from given datasets. It is capable of transforming quantitative values in datasets into linguistic terms, then filtering them and finding association rules. Each item uses only the linguistic term, thus making the number of fuzzy regions to be processed the same as that of the original items. The hybrid technique therefore focuses on the most important linguistic terms for better accurate output.

II. PROJECT SCOPE

The volume of information, in fields like medicine, business or educational institutes is massive and on far side human capability to extract valuable information. Apriori algorithm along with fuzzy logic (i.e. Hybrid technique) provides a way to obtain interesting patterns from datasets and also we are comparing their accuracy with the result obtained by applying Apriori technique alone. It also eliminates human error and provides high accuracy of results. Experiments will be conducted on Pima Indian Diabetes data set and Lung cancer dataset. The system is being used more and more frequently in intelligent systems because of its simplicity and similarities in human reasoning.

III. METHODOLOGY

In the proposed method we are dealing with incomplete dataset. For proposed methodology some necessary concepts are going to be used like fuzzy technique, Apriori algorithm, classification technique, Fuzzy rough set theory. In order to complete the incomplete datasets it is necessary to use rough set theory concept [3]. By applying fuzzy rough set theory with incomplete datasets retain more original dataset. After getting complete data sets, we proposed data mining approach that integrate fuzzy logic concepts with Apriori algorithm (Hybrid Technique) to find interesting item sets. In detail first quantitative datasets are converted into linguistic terms by utilizing membership function, fuzzy table will be created. After that applying Apriori algorithm complete set of rules are generated. Also we are comparing their accuracy with the result obtained by applying Apriori technique alone. We are compromising on time to achieve complete and accurate rules.

IV. BASIC MATHEMATICAL MODEL

Relevant Mathematics and Technical Keywords:
The system S
= {I, O, P, D}……(i)
where,

Input I
= { Membership function , Minimum_Support, Minimum_Confidence}…….(ii)

Output O
= {Association Rules, Complete_Datasets}…….. (iii)

Process P
= {Fuzzy Technique, Apriori algorithm, fuzzy rough set theory}…….(iv)

Dataset D
= {Pima diabetes dataset, lung cancer dataset}……...(v)

A] Support:

The support of an item set X can be defined as the proportion of transactions in the data set which contain the item set.

Supp(X) = no. of transactions that contain the item set X / total no. of transactions…….....(vi)

Supp(X) = Probability(X)……..(vii)

Supp(X, Y) = Supp(X U Y) …….(viii)

B] Confidence:

Conf(x=>y)

= Supp(X U Y) / Supp(X)

= Probability (Y/ X) …….. (ix)

C] Membership Function:

Fuzzy set A

A= {(x, μ A(x)), where x ∈X}…… (x)

where A(x) is called the membership function for the fuzzy set A.
The membership function maps each element x with a value in the interval [0, 1].

The following diagram is the triangular membership function which is specified by three parameters {a, b, c}.

$$\mu_s(x,a,b,c) = \begin{cases} 0, & \text{if } x<a \\ (x-a)/(b-a), & \text{if } a\le x\le b \\ (c-x)/(c-b), & \text{if } b\le x\le c \\ 0, & \text{if } c<x \end{cases}$$

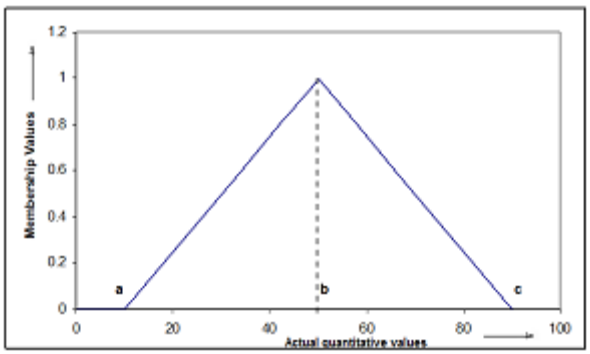

Figure 1: Triangular Membership_Function

V. Fuzzy Technique

Purpose: For simplicity and similarity to human reasoning in intelligence system.

Input: Incomplete quantitative data.

Processing: Using membership function to transform each quantitative value into a fuzzy set in Linguistic Terms.

Output: Large item set with candidate key.

VI. Apriori Algorithm

In Databases, the efficiency of mining association rules is an important field of Knowledge Discovery. The Apriori algorithm is a algorithm in mining association rules. Apriori employs an iterative approach known as level-wise search, where k-itemsets are used to explore (k+1)-itemsets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and gathering items that satisfy minimum support. The large 1-itemset is denoted as L1. Next, L2 is find using L1, the set of frequent 2-itemsets, which is used to find L3, and so on, till no more frequent k-itemsets can be found. Determining of each Lk needs one complete scan of the database.
At final iteration it will end up with many k-itemsets which is basically called association rules. Selecting interesting rules from the set of all possible rules various constraint measures such as support and confidence is applied.

Join step: meaning 1-itemset is made to self join with itself to generate 2-itemsets.

Prune step: here resulting set from join is filtered with minimum support threshold.

Cardinality set: resulting set from Prune step .

VII. J48 Decision Tree

A decision tree may be a prophetic machine-learning model that decides the target worth (dependent variable) of a brand new sample supported varied attribute values of the available data. The intermediate nodes of a decision tree denote the various attributes the branches between the nodes tell us the potential values that these attributes can have in the observed samples, whereas the terminal nodes tell us the ultimate value (classification) of the dependent variable.

The attribute that is to be predicted is known as the dependent variable, since its value depends upon the values of all the other attributes. The other attributes, which facilitate in predicting the value of the dependent variable, are known as the independent variables within the dataset.

The J48 Decision tree classifier has simple algorithm. To classify a new item, it first needs to create a decision tree based on the attribute values of the available training data. So, whenever it encounters a collection of items (training set) it identifies the attribute that discriminates the different instances most clearly. This character is able to tell us most about the data instances so that we can classify them the best is said to have the highest information gain. Now, among the possible values of this character, if there is other value for which there is no ambiguity, that is, for which the information instances falling inside its group have the same value for the final variable, then we terminate that branch and set to it the final value that we have obtained.

For the other cases, we then check for another attribute that gives us the best data gain. Hence we continue in this way until we either get a clear decision of what combination of attributes gives us a specific final target value. If we cannot get an unambiguous result from the available data, we set this branch a target value that the maximum items under this branch possess. After having the decision tree, we follow the sequence of attribute selection as we have obtained for the tree. Checking all the associated attributes and their values with those seen in the decision tree, we can set or predict the target value of this new instance.

VIII. Fuzzy Rough Set Theory

Fuzzy sets and rough sets address two important and mutually orthogonal, characteristics of incomplete data and knowledge. Rough set theory may be considered as the independent.[7] The main advantage of rough set theory in data analysis is that it does not need any preliminary or additional information about data.[7] While the former allow that objects belong to a set or relation to a given degree, the latter provide approximations of concepts in the presence of incomplete information. Fuzzy Rough set theory is a new mathematical approach to imperfect information [3]. The problem of imperfect information has been tackled for a long time by philosophers, logicians and mathematicians. Now it became also a crucial issue for computer scientists, there square measure several approaches to the problem of how to understand and manipulate incomplete or imperfect information [3]. The foremost productive one is, no doubt, the, the fuzzy rough set theory. In this paper, we demonstrate how these terms can be combined into a hybrid theory that is able to capture the best of knowledge. In particular, we review various alternatives for defining lower and upper approximations of a fuzzy set under a fuzzy relation, and also explore their application in query refinement.

IX. System Architecture

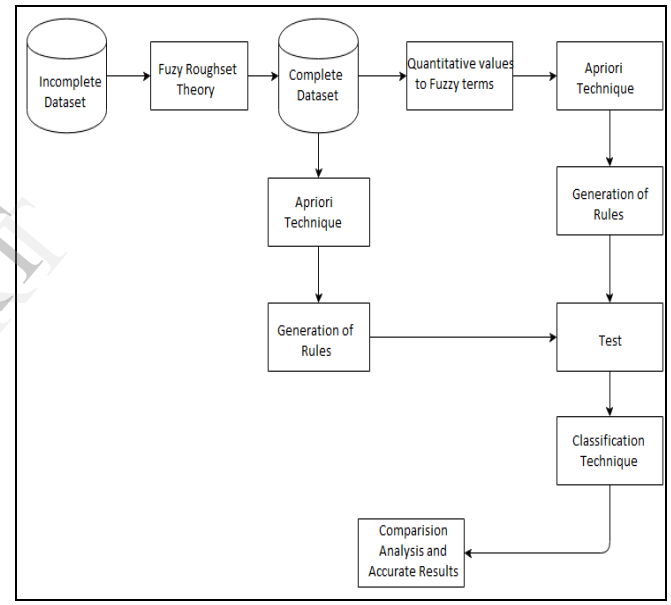

Figure 2: System Architecture

Above architecture shows actual flow of the system.

In recent days, machine learning and important knowledge searching techniques have attracted a good deal of attention in the information field area. Classification is also one of the prime research topics on this research field. Most of the explorations on classification concern that a complete data set is given as a training set and the test data know all values of attributes clearly. Miserably, incomplete information data are frequently seen in real-world applications. In this paper, we tend to propose a method to influence the incomplete quantitative data. Incomplete quantitative dataset is given as an input to the system. Fuzzy rough set theory is a pure mathematics may be a new mathematical approach to imperfect information. Applying fuzzy rough set theory on incomplete datasets to get complete data sets. Further hybrid technique (Fuzzy Technique along with Apriori technique) and Apriori alone is applied on complete datasets. Checking

the accuracy of the technique can be done by completeness on obtained rules.

X. CONCLUSION

In this paper, we have proposed generalized data mining algorithm to find interesting patterns among them although the proposed algorithm requires much computation time to derive more complete set of rules. Trade off thus exist between total computational time required and the complete set of rules. Choosing an appropriate domain the proposed algorithm can solve conventional transaction data problems by using degraded membership functions.

REFERENCES

[1] An object parameter approach to predicting unknown data in incomplete fuzzy soft sets by Tingquan Deng, Xiaofei Wang, Science direct ,2013.
[2] R. Agrawal and R. Srikant Algorithms for Mining Association Rules in Large Databases," Proc. 20th Int'l Conf. Very Large Data Bases, pp. 487-499, 1994.
[3] Mining from incomplete data by fuzzy rough sets by Tzung-Pei Hong a, Li-Huei Tseng b, Been-Chian Chien, Science direct ,2010.
[4] Knowledge and intelligent computing system in medicine by Babita Pandey, R.B.Mishra , Science direct ,2010.
[5] Mining frequent patterns and association rules using similarities by Ansel Y. Rodrguez-Gonzlez A, Jos Fco. Martnez-Trinidad A, Jess A. Carrasco-Ochoa A,Jos Ruiz-Shulcloper B, Science direct ,2013.
[6] Trade off Between computation time and no of rules for fuzzy mining from quantitative data.by Thung-PEI,Chan-Sheng,Sheng-Chai Chi,international journal uncertainty,2001.
7] Rough Sets by Zdzisław Pawlak, Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, ul. Bałtycka 5, 44 100 Gliwice, Poland.