On the Use of Vague Set Theory and Genetic Algorithm for Hesitation Information Mining

Prateek Shrivastava[1], Akhilesh Tiwari[2]
[1, 2] Department of CSE & IT
[1, 2] Madhav Institute of Technology and Science, Gwalior (M.P.), India

Abstract- *Existing methods utilizes the concept of association rule mining for finding frequent itemsets has been extensively used to discover interesting rules or relationships between items in large databases but it has limitations that it solely deals with the items or products that are sold but avoids the items that are nearly sold. These nearly sold things carry hesitation data since customers are indecisive to shop for them. This work proposed vague set theory that is capable of handling hesitation information of items. This paper describes that hesitation information of items is precious knowledge for the design of profitable selling strategies. This work proposed Genetic Algorithm based on evolution principles that has found its strong base in mining or maximize the rules for the items that customers mostly hesitate to purchase or has a high percentage of hesitation because of some reasons like price of an item, quality of an item, etc. Fitness function, crossover, and mutation are the main parameters involved in Genetic Algorithm which we used in our work. This work describes that if the reason of giving up the items is identified and resolved, we can easily remove this hesitation status of a customer and considering newly evolved rules as the interesting ones for boosting the sales of the item.*

Keywords:- Data mining, Association rule mining, Vague Sets, Hesitation Information, Genetic Algorithm.

I. INTRODUCTION

Data mining is a technique that helps to extract important data from a large database. It is the process of sorting through large amounts of data and picking out relevant information through the use of certain sophisticated algorithms. As more data is gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform this data into information.

Data mining is a process of extracting interesting patterns and valuable knowledge and information from huge volume of data. There are several techniques that have been used to discover such kind of specific and richer information, most of them resulting from statistics and machine learning[1].The jobs performed in the data mining depend on what kind of precious information someone needs to mine. Association Rule Mining is a characteristic of data mining which is used to observes relationships between things that are mostly buying together,[2]. Association rule mining determines interesting valid rules among a large data set of items.

This work describes how to mine useful valid rules from hesitation information using genetic algorithm that capture more richer information than traditional ARs. To describe the information of a customer's interest on items, we use the concepts of median membership and imprecision membership using vague set theory.

II. ASSOCIATION RULE MINING

Association rule mining is an important data mining model studied extensively by the database and data mining community. Initially used for Market Basket Analysis to find how items purchased by customers are related. In market basket data, consists of a large number of records and in each record all items bought by a customer on a single purchase transaction are listed. The interesting point would be paying attention to know that which groups of items are constantly purchased together. By applying association rule mining, we can get interesting relationships between distinct objects. Thus association rule mining becomes very interesting concept in dataset analysis.

Association rule generation is generally divided into two step process:

- To find all frequent item-sets in a database or to form a pattern, a user specified minimum support threshold is used.
- In order to form valid rules in large databases, a user specified minimum confidence threshold is applied to these frequent item-sets.

$$\boldsymbol{Support}(A \rightarrow B) = \frac{\text{No. of transactions containing both A and B}}{\text{Total no. of transactions}} \qquad (1)$$

$$\boldsymbol{Confidence}(A \rightarrow B) = \frac{\text{No. of transactions containing both A and B}}{\text{No. of transactions containing A}} \quad (2)$$

Existing association rule mining algorithms does not consider the hesitation information of an item, it only corresponds to support (buying) and against (not buying) information about item. Hesitation information of an items means customers are always hesitated to buy them because of some reasons like quality of an item, color availability of an item, price of an item, etc. In scenario of supermarket, hesitation information is precious knowledge for good selling strategies.

In our work, we concentrated on 'Association Rule Mining' technique for mining information from a transactional database. We implemented Apriori algorithm to generate association rules from a given database and then applying genetic algorithm for finding new interesting hesitation rules.

A. Apriori Algorithm

Apriori algorithm [3] was proposed by Agrawal and Srikant in 1994. The algorithm finds the frequent set L in the database D. Apriori is the improvement over association rule mining. It is one of the most commonly used algorithms that generate valid association rules. Traditional Apriori algorithm works on Boolean logic. Because of this reason Apriori algorithm has several drawbacks. It works on only yes (1) and no (0) form. It is inappropriate to handle imprecise and inexact data. It makes use of the downward closure property. The algorithm is a bottom search, moving upward level; it prunes many of the sets which are unlikely to be frequent sets, thus saving any extra efforts. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation, and groups of candidates are tested against the data). The algorithm terminates when no further successful extensions are found. Apriori uses breadth first search and a hash tree structure to count candidate item sets efficiently.

III. VAGUE SETS

In the real world there are vaguely specified data values in many applications, such as sensor information. Fuzzy set principle has been proposed to address such vagueness with the aid of generalizing the notion of membership in a set. Essentially, in a Fuzzy Set (FS) every detail of an element is associated with a point-value selected from the unit interval [0,1], which is called the grade of membership in the set. A vague Set (VS), as well as an Intuitionistic Fuzzy Set (IFS), is a in addition generalization of an FS. Rather than using point-based membership as in FSs, interval-based membership is used in a vague sets (VS). The interval-based membership value in VSs is greater expressive in capturing vagueness of records than point–based membership values. Fuzzy set idea has lengthy been introduced to deal with inexact and vague facts by using Zadeh's seminal paper in [4], seeing that within the actual international there's vague records approximately distinctive applications, together with in sensor databases, we will formalize the measurements from extraordinary sensors to a vague set. In fuzzy set concept, every item u ∈ U is assigned a single actual price, known as the grade of membership, among zero and one. (here U is a classical set of items, called the universe of discourse.). In [5], Gau et al. factor out that the disadvantage of using point-based value in fuzzy set concept is that the proof for u ∈ U and the proof against u ∈ U are in reality combined together. With the intention to tackle this problem, Gau et al. propose the perception of vague sets (VSs), which allow the use of interval-based membership instead of the use of point-based values as in FSs. The interval-based club generalization in VSs is extra expressive in capturing vagueness of facts. For that reason, the thrilling functions for handling indistinct information which are unique to VSs are largely left out.

Let I be a classical set of items, called the universe of discourse, in which detail of I is denoted by means of x.

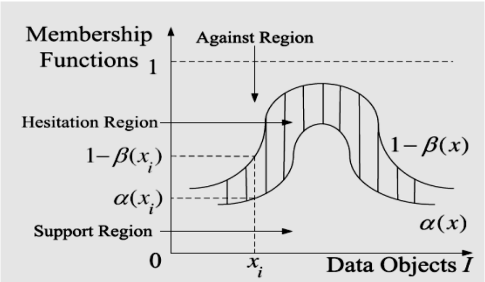

Figure 1: True (α) and False (β) Membership Functions of a Vague Set.

In this section, we discuss the following relationships of vague membership values (vague values for short) in VSs: crisp and imprecision. Remarkably, there are no such meaningful relationships based on IFS membership values. We also show two lattices arising from the crisp and the imprecision orders. In order to compare vague values, we need to introduce two derived memberships for discussion.

- The first is called the median membership, $Mm = (\alpha(x) + (1 - \beta(x)))/2$, which represents the overall evidence contained in a vague value. In addition, the vague value [1,1] has the highest Mm, which means the corresponding object totally belongs to the VS (i.e. a crisp value). While the vague value [0,0] has the lowest Mm which means that the corresponding object totally does not belong to the VS (i.e. the empty vague value).

- The second is called the imprecision membership, $Mi = ((1 - \beta(x)) - \alpha(x))$, which represents the overall imprecision of a vague value and is shown. In addition, the vague value [a, a], ($a \in [0, 1]$) has the lowest Mi which means that we know exactly the membership of the corresponding object (i.e. a fuzzy value). While the vague value [0,1] has the highest Mi which means that we know nothing about the membership of corresponding object.

Definition (Intent, Attractiveness and Hesitation, AH-Pair Transactions): The intent of an item x, denoted as intent(x), is a vague value $[\alpha(x), 1 - \beta(x)]$. The attractiveness of x, denoted as MA(x), is defined as the median membership of x, i.e., $MA(x) = (\alpha(x) + (1 - \beta(x)))/2$.The hesitation of x, denoted as MH(x), is defined as the imprecision membership of x, i.e., $MH(x) = ((1 - \beta(x)) - \alpha(x))$. The pair< _MA(x), MH(x)_> is called the AH-pair of x.

An object with high attractiveness approach that the item is nicely bought and has a excessive opportunity to be bought again next time. An item with high hesitation way that clients are constantly hesitating to shop for the item due to a few purpose (e.g., the client is anticipating rate reduction.

IV. GENETIC ALGORITHM

GA is an evolutionary algorithm which is proposed by Holland in 1975 [6]. They also trail the main beliefs of Charles Darwin Theory of endurance of the fittest. Genetic algorithm has been observed as a function optimizer. Genetic algorithms commence by initializing an inhabitants of solution which are known as chromosomes. It encompasses illustration of the problem typically in the bit vector form.

A. Some functions of genetic operators

(a) Selection: Selection deals with the probabilistic survival of the fittest, in that, greater in shape chromosomes are selected to continue to exist. Where in health is a comparable measure of how properly a chromosome solves the problem handy.

(b) Crossover: This operation is accomplished with the aid of deciding on a random gene alongside the length of the chromosomes and swapping all the genes after that factor.

(c) Mutation: Alters the brand new solutions so as to add inside the look for higher solutions. This is the risk that a chunk within a chromosome might be flipped (zero becomes 1, 1 becomes zero).

In this work for all generations, the fitness of each individual in the population is evaluated, multiple individuals are selected from the present population (based on their fitness), and modified (recombined and possibly mutated) to form a new population. Fitness function was defined as follows:

$$Fitness\,value = 1/(\alpha 1 * r1 + \alpha 2 * r2 + \alpha 3 * r3 + \alpha 4 * r4 + \alpha 5 * r5 + \alpha 6 * r6)$$

Where rn represents the value of the rules present in the database and coefficients α control effect of each parameter inside fitness function. Here, the values of α are fixed. For eg. α1 is 0.15. The selection of the fittest individual is done with the help of various methods.

V. RELATED WORK

A Lu, Yiping Ke, James Cheng, and Wilfred Ng[7] implemented vague set concept inside the context of AR mining as to encompass the hesitation facts into the ARs. Describe the concepts of attractiveness and hesitation of an item, which characterize the overall information of a customer’s intent on an item. Depending on these two concepts, proposed the notion of Vague Association Rules (VARs) and designed an effective algorithm to mine the VARs. Experiments demonstrate that the algorithm was effective and the VARs capture more exact and better information in turn to conventional ARs.

Anjana Pandey and K.R.Pardasani [8] presented a vague association rule to make available hesitation information and expand an algorithm to mine the hesitation information. The algorithm was devised to mine the courses and the hesitation of students to attend the courses. Experiments on real datasets confirmed that the algorithm to mine the Vague Association Rule is effective. In contrast to the conventional Association Rule mined from transactional databases, the Vague

Association Rule mined from the AH-pair databases are more detailed and capable to capture better information.

Bala Yesu Chilakalapudi, Narayana Satyala and Satyanarayana Menda [9] presented an algorithm for a resolving the issue problem of extracting frequent item sets from a huge vague database, interpreted under the Possible World Semantics (PWS). This issue is strictly difficult since a vague database consists of an exponential number of possible worlds. By examining the mining process can be modeled as a Poisson binomial distribution, an algorithm is implemented which can effectively and exactly determine frequent item sets in a huge vague database. The devised mining algorithm facilitates Probabilistic Frequent Item set (PFI) outcomes to be re-energized. The devised algorithm can maintain incremental mining and provides the precise outcomes on mining the vague database. The broad estimation on real data set to certify the scheme is performed.

VI. PROPOSED WORK

Traditional Association rule (AR) mining has a drawback that it only takes the items that are sold by the customer but ignores the items or products that are almost sold. In this work, we have proposed genetic algorithm for mining hesitation information of those items that customers are mostly hesitating to buy, by using vague set theory. If the reason of giving up the items is identified and resolved, we can easily remove this hesitation status of a customer and considering new evolved rules as the interesting ones for the design of profitable selling strategies.

We are using two algorithms, Cal_intent ( ) and Cal_AH-pair (intent) in our proposed algorithm which is described as:

Algorithm 1: Cal_intent ( )

```
1.  Initialize the intent array to store the intent.
2.  For each i = 0,1,2,3…..where i<no. of iid, do
3.  Initialize favor(α) and against(β) variable with value zero;
4.  For each j=0,1,2,3…….where j<no. of tid, do
5.  Increment favor(α) by one when D[i][j] is equal to one;
6.  Increment against(β) by one when D[i][j] is equal to zero;
7.  End;
8.  Generate intent as [α,1-β];
9.  End;
10. Return all intent;
```

This algorithm is iterative procedure to calculate intent which is first module of framework. This algorithm takes a dataset (D) as input. This dataset consists of rows and column as transaction ID (tid) and item ID (iid) of a supermarket. This algorithm returns an intent array.

The second module of framework is to calculate the AH-pair. The Cal-AH-pair Algorithm is a simple iterative method that calculates AH pair. This algorithm takes intent as input which is calculated in previous algorithm.

Algorithm 2:Cal_AH-pair (intent)

```
1. Initialize an array AH pair;
2. For each i=0,1,2,3…...where i< no. of iid
3. Attractiveness as a median membership i.e.
   ½(α+(1-β));
4. Hesitation as a difference of α and 1-β
   using intent;
5. End;
6. Return all AH-pair
```

We analyzed that hesitation for an item decrease the attractiveness of an item and hence reduce the probability of selling the item. This reduces the profitability of store in many folds. Here we analyzed the case of supermarket and find the main cause of hesitation in customer’s point of view. This algorithm is efficient in increasing selling probability of an item in case of supermarket scenario.

Proposed Algorithm:

Input: Hesitation database D, min support, min confidence.

Output: Profitable valid rules.

1. First initiate the hesitation database D that contains the values such as yes, no, and hesitation of selling of products.
2. Applying vague set theory to find vague values and AH-pair values from algorithm Cal_intent ( ) and Cal_AH-pair (intent) respectively, which indicates hesitation percentage of each items.
3. Then applying Apriori algorithm on hesitation database D to generate valid rules, before it taking the value of yes, no, and hesitation in the form of 0 and 1 i.e. called binary conversion; Y=1, H≅0, and N=0.
4. Now applying genetic algorithm on generated rules.

(a) First, find fitness function of each generated rules by the formula:-
FitnV =1/ (α1* r1 + α2*r2 + α3*r3 + α4*r4 + α5*r5 + α6*r6)
(b) Selection is done according to fitness values of rules.
(c) Single point crossover is applied.
(d) Single bit mutation is applied.
(e) New evolved hesitation rules.
(f) Repeat steps from (a) to (e) until termination condition is reached (FitnV < 0.9527).

5. End.

In this work, we take hesitation database as input that contains three types of information about items in the form of Y, N, and H. Y=yes (favor), N= no (against), H= hesitation, according to vague set theory which is capable of dealing with vague situations. Favor means customer is interested to buy an item. Against means customer is not interested to buy. Hesitation information means customer is always hesitated to buy.

After that vague value or intent value of an items is calculated. These values shows the favor(support), against, and hesitation percentage of each items. Then calculate attractiveness-hesitation(AH) values of each items.

An object with high attractiveness approach that the item is nicely bought and has a excessive opportunity to be bought again next time. An item with high hesitation way that clients are constantly hesitating to shop for the item due to a few purpose (e.g., the client is anticipating rate reduction).

Then applying Apriori algorithm on hesitation database to generate valid rules by taking user specified minimum support threshold and minimum confidence threshold. Minimum support is used to find frequent patterns. While minimum confidence is used to find valid association rules.

Applying genetic algorithm on rules generated from Apriori algorithm gives additional rules to be considered which is called as hesitation rules. Fitness function, crossover, and mutation are the main parameters involved in Genetic Algorithm which is used in this work.

This work shows that if the reason of giving up the items is identified and resolved, seller can easily remove this hesitation status of a customer and considering newly evolved hesitation rules as the interesting ones for boosting the sales of the item.

VII. CONCLUSION

In this work, we have presented a novel genetic based algorithm to mine hesitation association rules in big data sets. We apply the vague set based approach to model the hesitation information of the items. This work gives more valid rules i.e. converting the hesitation of higher order into attractiveness and thus giving user an idea about the possibility of the rules to be generated. This concept helps in exposing necessary hesitation rules that may also be considered for profitable decision making process. The performance of the proposed algorithm in terms of finding interesting hesitation rules gives more optimized results as compared to existing ones.

REFERENCES

[1] J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, 2nd ed., The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor (2006).

[2] Frawley, William J.; Piatetsky-Shapiro, Gregory; Matheus, Christopher J.: “Knowledge Discovery in Databases”: an Overview. AAAI/MIT Press, (1992).

[3] Rakesh Agrawal and Ramakrishnan Srikant, “Fast algorithms for mining association rules in large databases”, In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, Proc. of the 20th International Conference on Very Large Data Bases, pp. 487-499, (1994).

[4] L. A. Zadeh, fuzzy sets inf. Control 8 (1965), pp. 338-353.

[5] W. L. Gau and D. J. Buehrer. Vague sets. IEEE Transactions on Systems, Man and Cybernetics, 23:610–614, (1993).

[6] J.H. Holland, "Genetic algorithms and the optimal allocation of trials, SIAM J. Computing, pp. 88–105, (1973).

[7] An Lu, Yiping Ke, James Cheng, and Wilfred Ng, “Mining Vague Association Rules”, Department of Computer Science and Engineering The Hong Kong University of Science and Technology Hong Kong, China.

[8] Anjana Pandey and K.R.Pardasani, "A Model for Mining Course Information using Vague Association Rule", International Journal of Computer Applications (0975 – 8887) Volume 58– No.20, November (2012).

[9] Bala Yesu Chilakalapudi, Narayana Satyala and Satyanarayana Menda, "An Improved Algorithm for Efficient Mining of Frequent Item Sets on Large Uncertain Databases", International Journal of Computer Applications (0975 – 8887) Volume 73– No.12, July (2013).