Vague Set Theory for Profit Pattern and Decision Making in Uncertain Data

Vivek Badhe
Department of Computer Applications
MANIT, Bhopal, India

Dr. R.S Thakur
Department of Computer Applications
MANIT, Bhopal, India

Dr. G.S Thakur
Department of Computer Applications
MANIT, Bhopal, India

***Abstract*—Problem of decision making, especially in financial issues is a crucial task in every business. Profit Pattern mining hit the target but this job is found very difficult when it is depends on the imprecise and vague environment, which is frequent in recent years. The concept of vague association rule is novel way to address this difficulty. Merely few researches have been carried out in association rule mining using vague set theory. The general approaches to association rule mining focus on inducting rule by using correlation among data and finding frequent occurring patterns. In the past years *data mining* technology follows traditional approach that offers only statistical analysis and discovers rules. The main technique uses support and confidence measures for generating rules. But since the data have become more complex today, it's a requisite to find solution that deals with such problems. There are certain constructive approaches that have already reform the ARM. In this paper, we apply concept of vague set theory and related properties for profit patterns and its application to the commercial management to deal with Business decision making problem.**

Keywords*—Association Rule Mining; Vague Association Rule Mining; Profit Pattern Mining

I. INTRODUCTION

Pattern discovery from huge volume of data is one of the most desired attributes of Data Mining [5]. However, in reality, a substantial portion of the available information is stored in text databases, which consists of large collections of documents from various sources, such as news articles, books, digital libraries and Web pages. Since web search engines have become pervasive and search has become integrated, retrieving of information from these search engines consist of three essentials: query, documents, and search results.

The emerging growth of data mining raises the large range of complex applications [14]. It leads the broad study of data mining frequent patterns. Mining frequent sets over data streams present attractive new challenges over traditional mining in static databases. Data mining is generally used for retrieving the desire information to make it into knowledge from the large size databases.

The study shows [16] that interestingness measures are distinct for different applications and substantiate that domain knowledge is necessary to the selection of an appropriate interestingness measure for a particular assignment and business objective and the goal of any business is to generate profit. So the profit can be taken as one of the measures with proper mining technique so it can help in decision making process of business.

The rest of the paper is organized as follows, in section 2 we discuss the fundamental basis of association rules, and vague set theory that deal with uncertainty and vagueness, Section 3 introduces the related work that has been done so far with this theory in accordance with association rules called Vague Association Rule (VAR)[10]. In Section 4 describe our methodology and discuss the How VAR is helpful to business problem and its decision making process. In section 5 discussions of result and comparison with classical and vague rules. Finally Section 6 concludes the paper.

II. PRELIMINARIES

A. Association Rule Mining

Association rules discovery is one of the most important method which was given by R. Agrawal in 1993 [1]. It gives the information like "if-then" statements. These rules are invoked from the dataset. It generates from calculation of the support and confidence of each rule that can show the frequency of occurrence of a given rule. Association Analysis is the process of discovering hidden pattern or condition that occurs frequently together in a given dataset. Association Rule mining techniques looks for interesting associations and correlations among data set. An association rule is a rule, which entails probabilistic relationship, with the form $X \Rightarrow Y$ between sets of database attributes, where X and Y are sets of items, and $X \cap Y = \phi$. Given the set of transactions T, we are interested in generating all rules that satisfy certain constraints. These constrains are *support* and *confidence*. The *support* of the rule is the fraction of the transactions in T that satisfy the union of items in X and Y. The probability, measured as the fraction of the transactions containing X also containing Y, is called the *confidence* of the rule.

Support should not be confused with confidence. While confidence is a measure of the rule's strength, support corresponds to statistical significance.

With the help of these constraints, rules are computed from the data and, association rules are calculated with help of probability. Mining frequent itemsets [17] is a fundamental and essential problem in many data mining applications such as the discovery of association rules, strong rules, correlations, multi-dimensional patterns, and many other important discovery tasks. The first and foremost algorithm that was given to

generate association rules was *apriori* [2]. Its proposal used the same two constraints: support and confidence, and forming rules in accordance with these constraints.

B. Profit Pattern Mining

Ke Wang, Senqiang Zhou, and Jiawei Han in 2002 presented a profit mining [18] approach to reduce the gap between the statistic-based pattern extraction and the value-based decision making. They took a set of past transactions and pre-selected target items, and intended to build a model for recommending target items and promotion strategies to new customers, with the goal of maximizing the net profit. They identified several issues in profit mining and proposed solutions. They evaluate the effectiveness of this approach using data sets of a wide range of characteristics. The key to profit mining is to recommend "right" items and "right" prices. If the price is too high, the customer will go away without generating any profit; if the price is too low or if the item is not profitable, the profit will not be maximized. The approach is to exploit data mining to extract the patterns for right items and right prices. The key issues in this context are Profit based patterns, shopping on unavailability, explosive search space, optimality of recommendations, and interpretability of recommendation.

C. Vague Set Theory

The classical (crisp) set theory define sets as the "collection of objects (either similar or dissimilar) called elements of a set as a whole". These are also referred as crisp in nature because they only tells whether an element is a member or not, i.e., either 0 or 1. It may be also given as an element belongs to or does not belong to particular set. The crisp set theory often times is unable to provide a better understanding of any object/element to be of a certain group. Thus, leading to the fact that value might lie in between 0 and 1.

A *vague* set is a set of element distributed in a universe that has a grade of membership values in the continuous subinterval of [0, 1]. Hence, such a set can be marked by *true membership* and *false membership* functions. The continuous subinterval states both about the evidence that is in favor of the object and also that is opposing it.

In early 90's Gau's and Buehrer [4] introduced the notion of vague sets. Let V be the vague set. If U is the universe of discourse having X objects with *x* elements than V in U can be defined using the true membership (V_t) and false membership (V_f) functions. Considering that both V_t and V_f consorted as real numbers in the subinterval of [0, 1]. Also V_t is the lower bound on grade of membership of *x* derived in favor of *x*, and V_f is the lower bound on grade of membership derived against *x*, with each element in X where $V_t + V_f \leq 1$ and V_t:X→[0,1], V_f:X→[0, 1]. Hence, the grade of membership of *x* is bounded to a subinterval $[V_t(x), 1\text{-}V_f(x)]$ of [0, 1]. A vague set V in a universe of discourse U is characterized by a true membership function, t_v and a false membership function, f_v, as follows:

$$t_v: U \rightarrow [0, 1],$$

$$f_v: U \rightarrow [0, 1], \text{ and}$$

$$t_v(x) + f_v(x) \leq 1,$$

where $t_v(x)$ is a *lower bound* on the grade of membership of u derived from the evidence for *x*, and $f_v(x)$ is a lower bound on negation of *x* derived from the evidence against *x*. An Lu and Wilfred Ng [9] gave a detailed discussion of using the proper theory for imprecise or vague data.

D. Vague Association Rules

The notion of *Vague Association Rules* (*VARs*)[10] is based on four types of support and confidence, which applied on AH-Pair Transactions Given the transactions of the customers, we then aggregate the transactions to obtain the *intent* of each item. Based on the intent of an item, we next define the *attractiveness* and *hesitation* of it [6,7,8] .

(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 M_A (x), is defined as the median membership of x, i.e., $M_A(x) = (\alpha(x) + (1 - \beta(x)))/2$. The hesitation of x, denoted as M_H (x), is de-fined as the imprecision membership of x, i.e., $M_H(x) = ((1 - \beta(x)) - \alpha(x))$. The pair $hM_A(x), M_H(x)i$ is called the AH-pair of x. An AH-pair transaction T is a tu-ple $<v_1, v_2, \ldots, v_m>$ on an itemset $I_T = \{x_1, x_2, \ldots, x_m\}$, where $I_T \subseteq I$ and $v_j = hM_A(x_j), M_H(x_j)i$ is an AH-pair of the item x_j, for $1 \leq j \leq m$. An AH-pair database is a sequence of AH-pair transactions.

(Vague Association Rule) *A* Vague Association Rule (VAR), r = (X => Y), *is an association rule obtained from an AH-pair database.*

Based on the attractiveness and hesitation of an item, we define four different types of support and confidence of a VAR depending on what kind of knowledge we want to acquire. For clarity, we use A to denote *Attractiveness* and H to denote *Hesitation*.

Support and Confidence

Given an AH- pair database, *D* , we can define different types of support and Confidence for an itemset *Z* or a VAR $_X \Rightarrow _Y$,where $X \cup Y = Z$ [6].

Support $T_S = \sum Mp / |D|$

Where

T_S = {A-sup, H-Supp,AH-Supp,HA-Supp}
p = {X,Y,XUY}
M = $\{M_A, M_H, M_A.M_H\ M_H M_A\}$

Confidence C = $T_S(Z)/ T_S(X)$

E. Uncertainty in Data Mining

The traditional data mining approach uses statistical and logical significance to find the knowledge from the databases. Since the databases have become diverse and heterogeneous which contain data close to real world, they are susceptible to uncertainty. By uncertainty we mean that it is not possible to depict the true nature of the data and what will be the outcome of it when processed. Uncertainty occurs when it is impossible to assert any value to an object when modeling is done [9,10]. Uncertainty can be of distinct forms and to identify specific one is a taxing work. Some types of uncertainties are:

- *Imprecision*: the available information is not specific to the desired modeling.
- *Inconsistency*: there are two or more statements in modeling which cannot be true at same instant.
- *Ambiguity*: the objects in the model have stringency because of which many possible renditions can be made.
- *Vagueness*: the objects in a model include an intrinsic vague value which is not expressed clearly. Vagueness is formalized from the concept of fuzziness.

To deal with uncertainty in mining, some soft computing techniques must be incorporated which helps to reason with the databases. Neural Networks, Fuzzy logic, Genetic Algorithm, Rough Sets, Vague Sets are some of the soft computing techniques that does deal uncertainty to some extent].

III. Related Work

In ARM, support and confidence are the basic measures that have been used since its inception, which define the statistical significance of any rule [16].

Sandhu, P.S. et. al. in 2010 [15] proposed an efficient approach based on weight factor and utility for effectual mining of significant association rules. Initially, the proposed approach makes use of the traditional Apriori algorithm to generate a set of association rules from a database. The proposed approach exploits the anti-monotone property of the Apriori algorithm, which states that for a k-itemset to be frequent all (k-1) subsets of this itemset also have to be frequent. Subsequently, the set of association rules mined are subjected to weight age (W-gain) and utility (U-gain) constraints, and for every association rule mined, a combined Utility Weighted Score (UW-Score) is computed. Ultimately, they determined a subset of valuable association rules based on the UW-Score computed. The experimental results demonstrate the effectiveness of the proposed approach in generating high utility association rules that can be lucratively applied for business development

An Lu and Wilfred Ng [10] provided another 4 support and 4 confidence measures based on vague properties which assists in finding more interesting rules. Merely few researches have been carried out in association rule mining using vague set theory [3].

In 2007, An Lu and Wilfred Ng[10] apply the vague set theory to address a limitation in traditional AR mining problem, that is, the hesitation information of items is not considered. They propose the notion of VARs that incorporates the hesitation information of items into ARs. They also define different types of support and confidence for VARs in order to evaluate the quality of the VARs for different purposes. An efficient algorithm is proposed to mine the VARs.

An Lu et al. [11] Modeled hesitation information by the purchaser in online shopping using Vague Association rule and provide the notion of VAR for almost sold items in the online shopping by considering different user preference.

Anjna Pandey et al. [12,13] developed the models for hesitation information for course information using vague set theory in order to address a limitation in traditional association rule mining problem, which ignores the hesitation information of items in transactions. The efficient algorithm for mining vague association rule that discovers the hesitation information of items is proposed to solve the course information, attendance and related vagueness. They extend their work for temporal Association rule mining that can be used to evaluate the course effectiveness and helps to look for in regards to changes in performance of the course from time to time.

IV. Proposed Methodology

The proposed methodology is especially developed for commercial transactions where each item having an item code but for the different packing the code is differ for the same item. This will create the vagueness and cannot be deal by the conventional association rule mining for this purpose we use the vague association rule mining and generate those rules which generate the profit significance but ruled out due to statically violation caused by vagueness. We use an algorithm to mine Vague set based Association Rules. As discussed in previous section, the vague sets have found application in association rule mining in many ways. We propose another important methodology to find support and confidence for mining association rules. The technique we propose consist three new formulas for finding support and confidence.

Definition 1: Variation Table/Matrix

The variation table matrix is a table that is formed of vague items. The variation table contains N rows depicting the number of vague items and M columns corresponding to the variant of that vague item.

Definition 2: Vague Percentage

The vague percentage denotes the amount of vagueness contained in a database for a particular vague item. For an item in a database, there exists a true membership (V_t), a false membership (V_f), and a certain amount of vagueness, as in our case vague percentage (V_p). Thus, a database consists of $|D| = V_t + V_f + V_p \leq 1$ all the memberships and percentage values. The vague percentage can be denoted now as $V_p = |D| - (V_t + V_f)$.

Definition 3: True Support (Sup_{tr})

Let A be a vague item that has A_1, A_2… as vague values, then true support Sup_{tr} is defined as $Sup_{\mathrm{tr}} = \frac{V_t(A_1)}{[V_t(A_1)+V_f(A_1)]}$ i.e. the true membership of an item A_1 is in ratio with the sum of its own true membership and its false membership. We do this because the database as a whole contains vagueness, which means $|D| = V_t + V_f + V_p \leq 1$. Hence on excluding vagueness from the database $(V_t + V_f \leq 1 - V_p)$ then it gives the true support of that item which classical method did not consider. The value of Sup_{tr} can also be defined in terms of Vague Percentage as $\frac{\left\{V_t + \left[\frac{V_P * V_t}{(V_t + V_f)}\right]\right\}}{|D|}$

Definition 4: True Confidence ($Conf_{tr}$)

True confidence is the ratio of true support of the union of items A and B to the true supports of any of the item either A of B. It is denoted by $Conf_{tr} = \frac{Sup_{tr}(A \cup B)}{Sup_{tr}(A)}$.

Algorithm: Vague Itemset minEr (VIE)

A. Vague Table Generation Algorithm

Variation Table : VagueTab (D)

1) Scan the database D to find the total number of distinct items;

2) For i = 0, 1, 2, …where i = no. of transactions T in D, do

3) Initialize both true membership (V_t) *and false membership* (V_f) *variables with zero;*

4) For j = 0, 1, 2, … where j=no. of vague items in D, do

5) Increment the vague item count for i^{th} *item and store it in the Vague Table;*

6) End of for;

7) End of for;

8) Return VagueTab(t);

B. Vague Itemset minEr (VIE) Algorithm

VIE (**D**, V_t)

1) Calculate 1^{st} *vague frequent itemset by scanning D;*

2) For i=0, 1, 2, … t_n *transactions in D, do*

3) Call VagueTab(D) and generate candidate C_i *and check whether the item in candidate is vague or not;*

4) If an item is vague, find its variant from VagueTab and calculate true membership (V_t) *of that item(s);*

5) If an item is non-vague then directly add in to the candidate list C_i*;*

6) Perform the pruning of the list by applying true membership (V_t) *and generate* 2^{nd} *vague frequent itemset;*

7) Find next vague frequent itemset till no further combinations are possible

8) End of for;

9) Return vague frequent itemset (L);

C. Vague Rules Generation

VagueR(sup, conf)

1) find the last vague frequent itemset (L);

2) generate subsets, s of the vague itemset such that $s \in S$*;*

3) if s is vague frequent itemset, then find the rule;

4) else omit the subset from the itemset list;

5) return VagueR(n, m);

The experiment was conducted on an FMCG database that contains the daily inventory of the products purchased by users. We classify items on the basis of products denoting each with a certain unique code. The vagueness is found in the database by using the Vague Table which consist all the variants of a particular product that is available in the database. This imparts vagueness in the dataset on which we apply out Vague Itemset minEr (VIE) algorithm. Since the FMCG database is very large consisting of huge number of transactions, we only report results on a sample of transactions selected at random. The number of transactions *T* taken into account is 10 which consists a number of products with their codes, i.e. both vague items and non-vague items. First experiment was conducted on the sample dataset with the traditional Apriori method and the results are noted in Table 1. The second experiment was conducted on the same sample dataset with the Vague Itemset minEr (VIE) algorithm and the result are noted in Table 2. The minimum threshold support was kept at 30% and minimum threshold confidence at 80%.

TABLE I. RESULT OF APRIORI

		(min_sup, min_conf)
AAT001	<- ABD022	(50, 80)
AAT001	<- ABB012	(50, 80)
AAE038	<- ABB012	(50, 80)
AAT001	<- AAE165 ABB012	(30, 100)
ABB012	<- AAE165 AAT001	(30, 100)
AAT001	<-ABD022 ABB012	(30, 100)

TABLE II. RESULT OF VAGUE CONSIDERED ARM

		(min_sup, min_conf)
ABD022	<- AAG272	(40, 82)
AAE038	<- AAE165	(40, 82)
ABB012	<- AAE165	(40, 82)
ABB012	<- AAT001	(60, 83)
AAE038	<- ABB012	(83, 96)
ABB012	<- AAE038	(85, 94)
ABB012	<- AAE165 AAT001	(30, 100)
AAT001	<- AAE165 ABB012	(33, 90)

It is clear from the Table 1 & 2 that the rules that are generated from traditional Apriori have only three rules (highlighted above for visual understanding) that are also found using VIE algorithm but with difference in their support and confidence measures. Some of the rules that are in Table 1 are omitted in Table 2 and vice versa. This demonstrates that the traditional Apriori performs undermining and forms *subversive rules* whereas the vague algorithm VIE performs over mining and forms *puissant rules*. Thus vague sets when incorporated with association rules provide a contrasting meaning to the classical approach and gives better results.

V. DISCUSSION

The classical approach to association rule mining uses the Apriori or similar algorithms which are based purely on statistical significance of the items present in the database. The approach was easy to consider when the databases contained only textual data, or in other words, certain data. As the database technology evolved, the basic measures of support and confidence were proving to be scarce in finding relevant knowledge. It is evident that without support and confidence it is difficult to find rules but to improve the knowledge discovery some more measures or dimensions need to be incorporated. One way of doing it is by first understanding our database. There are many mathematical tools that have been used over time to fulfill ones requirement with databases. The approach we propose takes uncertainty in consideration. Now, there are many types of uncertainty that could be handled each

with a different and specific tool. The uncertainty we consider is of vagueness type.

Vagueness is the property of an item that is difficult to comprehend and differentiate. The principles of vague set theory are used to deal with vagueness. Unlike fuzzy logic which is a special case of vague logic, the vague sets allows to bound the existence of item(s) to an interval. Any item belonging or not will be denoted by its true and false membership. We incorporate vague logic with the classical Apriori algorithm to find more relevant yet vague rules.

VI. CONCLUSION

Vague Association Rule Mining for profit pattern combine the statistic based pattern extraction with value-based decision making to achieve the commercial goals. The work we propose gives an advantage by incorporating vague sets in data mining. The vague sets allow us to consider the vague uncertainty existing in databases and to utilize it to mine such rules that ultimately give better correlation among items. The result calculated was better in comparison to the traditional approach and gives an alternative approach to data mining. The proposed approach not only improves the mining process but also provide the profitable rules in uncertain data. Although a many researches has been carried out in association rule mining but still it requires more attention for defining the notion of profit which would help in improving business strategies and provide some recommender rules.

ACKNOWLEDGMENT

This work is supported by research project under Fast Track Scheme for Young Scientist from DST, New Delhi, India. Scheme 2011-12, No. SR/FTP/ETA-121/ 2011 (SERB), dated 18/12/2012.

REFERENCES

[1] Agrawal Rakesh, Imielinski Tomas, Swami Arun, "Mining Fuzzy Weighted Association Rules" 1993 © SIGMOD ACM

[2] Agrawal Rakesh, Srikant Ramakrishnan, "Fast Algorithms for Mining Association Rules - A Priori", © 1994

[3] Badhe V, Thakur R.S., Thakur G.S., "A Review on Dealing Uncertainty, imprecision and Vagueness in Association Rule Mining Using Extended and Generalized Fuzzy" IJECS, Vol. 3, issue 7, 2014

[4] Gau Wen-Lung and Buehrer Daniel J., "Vague Sets", © 1993 IEEE

[5] Han J. and Kamber M., "Data Mining: Concepts and techniques", Morgan Kaufmann Publishers, Elsevier India, 2001.

[6] Lu, An., Ng,W "Managing merged data by vague functional dependencies". In: Atzeni, P., Chu, W., Lu, H., Zhou, S., Ling, T.-W. LNCS, vol. 3288, pp. 259–272. Springer, 2004

[7] Lu An and Ng Wilfred "Maintaining consistency of vague databases using data dependencies" Data and Knowledge Engineering, Volume 68,2009,Pages 622-641.

[8] Lu.A.,Ng.W:Handling Inconsistency of vague relations with functional dependencies. Springer 2007.

[9] Lu An and Ng Wilfred, "Vague Sets or Intuitionistic Fuzzy Sets for Handling Vague Data- Which One Is Better?" 2005 © Springer

[10] Lu An, Ke Yiping, Cheng James, and Ng Wilfred, "Mining Vague Association Rules" 2007 © Springer

[11] Lu An and Ng Wilfred "Mining Hesitation Information by Vague Association Rules"Lecture Notes in Computer Science ,Springer Volume 4801/,2008,pg 39-55.

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

[13] Pandey Anjana, Pardasani K.R. ," A Model for Vague Association Rule Mining in Temporal Databases, Journal of Information and Computing Science, Vol. 8, No. 1, 2013, *pp*. 063-074

[14] Pujari A. K., Data Mining Techniques, University Press 2001.

[15] Sandhu, P.S.; Dhaliwal, D.S.; Panda, S.N.; Bisht, A., "An Improvement in Apriori Algorithm Using Profit and Quantity" ICCNT Year: 2010, IEEE conference publication.

[16] Tew. C, Giraud-Carrier C, Tanner K, Burton S. "Behaviour based clustering and analysis of interesting measures for association rule mining" Springer 2013.

[17] Tiwari A., Gupta R.K. and Agrawal D.P. "A survey on Frequent Pattern Mining : Current Status and Challenging issues" Information Technology Journal 9(7) 1278-1293, 2010.

[18] Wang Ke, Zhou Senqiang, and Han Jiawei, Profit Mining: From Patterns to Actions, C.S. Jensen et al. (Eds.): EDBT 2002, LNCS 2287, pp. 70–87, 2002.Springer-VerlagBerlin.