



Incorporating Vague Set Theory for Decision Making Process in Association Rule Mining

Terrence Shebuel Arvind

M.Tech. Scholar, Gyan Ganga College of Technology,
Jabalpur (M.P), India

Vivek Badhe

Asst. Professor, Gyan Ganga College of Technology,
Jabalpur (M.P), India

Abstract: Yesteryear data mining technology follows traditional approach that offers only statistical analysis and discovers rules. In light of recent advancements in database technology, the data stored in databases are more “real world” and “unstructured” thus they are prone to have some amount of uncertainty. Dealing with uncertainty is a challenge faced in many research areas including data mining. There are many works that uses mathematical approach to deal uncertainty. To name a few are probability theory, fuzzy logic approach, rough set and soft set approach. All these theories deal with different kinds of uncertainties. Our work demonstrates the uncertainty of vagueness type and addresses how to handle vagueness in data mining, if one encountered, by using the principle of vague set theory and then employ it to find association rules.

Key Terms: Data Mining, Association Rule (ARs) Mining, Vague Set Theory, Subversive Rules, Puissant Rules

I. INTRODUCTION

In recent years, databases have evolved in many aspects. Not only can they store data in structured form but in unstructured too. The foremost process to analyze these databases to concur some useful information was called Knowledge Discovery in Databases (KDD). The purpose of KDD process is to mine hidden knowledge from large databases which might be both useful and fruitful. For this data mining emerged as an area for market-basket analysis where buying patterns of consumers were found interesting and new schemes and marketing was done accordingly. But the mining became pervasive and is now applicable to, if not all, many areas of research. There are a number of mining techniques being in use. The earliest of them all is the Association Rules Mining [2]. This method includes basic statistical analysis to find patterns of interest and later generate valid rules that will affect the outcome of its application. As the databases become more advance in storing data of heterogeneous nature, it was evident that statistical techniques will not provide an adequate result. Because of which many reasoning based techniques that could be incorporated with data mining started to surface up. To reason with a computer system is a critical task. Many machine learning and artificial intelligence systems are being build that does the work of humans. The human tasks are uncertain in nature and so is the data we generate by utilizing all the technology surrounding us. To mine this data in a way that ultimately results in decision making is complex. First what is needed is to identify the database for a particular type of uncertainty and then deal with it. Many mathematical tools have been developed and used so far like probability, fuzzy theory [16], rough sets [14], soft sets [11] etc. We use vague set theory [6] to find vagueness in the databases. The vague set theory allows us to consider those values, those objects that are often neglected from the process. The databases today contain more vagueness then they did a decade ago and in near future it is likely to increase in vague percentage. Thus, vague set theory is an important aspect that could be used with data mining.

II. PRELIMINARIES

A. Association Rule Mining

Association Rule Mining was first introduced by Rakesh Agrawal et al. in 1994 [2] as a research program to find frequent patterns in very large databases. Prior to it, the same publish a method known as AIS [1] which was the beginning in field of data mining. Later, *a priori* algorithm was proposed which is so simple and unique that it has been for the past two decades a milestone in association rule mining research. The association rules are deducted by applying frequent pattern mining where the occurrence (frequency) of an object (or item) is counted, thus it is used to mine frequent patterns that have some interestingness contained and later can be used to generate rules.

The association rule mining algorithm, *a priori*, is one of the most used, worked upon, enhanced algorithm that is easy to comprehend. The *a priori* takes the basic statistical approach by finding the occurrences of any item in a transaction in the database. Thus a frequent patterns and rules can be found just by simple counting of items. After these patterns are found, the correlation between the items of frequent patterns is determined. The

correlation tells about how an item will affect any other item in the process of rule generation. But to have a valid frequent pattern the a priori uses two measures known as *support* and *confidence*. *Support* is the union of the items that are present together in a transaction T in the database D . While confidence is the conditional probability that an item can be found to generate rule. These two measures are responsible for the working process of finding interesting patterns as well as rules.

In a database, let items X and Y are in correlation with one another then based on Apriori, the frequent patterns are discovered by counting the database for that particular item, either X or Y . The count of the item must satisfy some support threshold value. If it does then it is assumed as frequent and if it does not satisfy then it is called an infrequent item/itemset. Thus the support value can be defined as $Supp(X) = \frac{count(X)}{|D|}$, where $|D|$ is the database.

Similarly any item's support can be calculated by the formula. However, the rule generation requires minimum a 2-frequent itemset so that an actual rule can be established. The rule $X \Rightarrow Y$ is a valid rule if the values satisfy the confidence threshold. Thus a rule can be defined as the conditional probability $P(X|Y) = X \Rightarrow Y = \frac{Supp(X \cup Y)}{Supp(X)}$.

The validity of rule solely depends on the measures. For some reason the support is kept at lower value while confidence is kept high.

B. 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 [15, 12]. 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 and have found their use in the research field of association rule mining [3, 4, 5].

C. Vague Set Theory

Vague set theory [6] was introduced by Gau et al. in 1993. The theory is based on the concepts of fuzzy set theory but does provide a more accurate and precise view of the data. The fuzzy set theory allows us to define any object in the universe with a membership value. This membership value is of graded nature thus gives us an enhanced view of the model that was not possible with the classical set theory. Fuzzy set theory tells us about the true membership percentage of an object.

Vague set theory on the other hand not only tells about the true membership but false too. Thus more precise information is visible in vague sets rather than fuzzy sets.

Let U be the universe where X is a set of objects containing x elements. A vague set V can be defined as *true membership* (V_t) and *false membership* (V_f), where $V_t(x)$ is the graded membership lower bound of x derived from the evidence of x , and $V_f(x)$ is the graded membership lower bound of *negation* of x derived from the evidence *against* x . The true and false memberships $V_t(x)$ and $V_f(x)$ are confined within the interval $[0, 1]$ for all x in X , thus $V_t(x) + V_f(x) \leq 1$. Thus

$$\begin{aligned} V_t : X &\rightarrow [0, 1] \\ V_f : X &\rightarrow [0, 1] \end{aligned}$$

bounds the grade of membership of x to the subinterval $[V_t(x), 1 - V_f(x)]$ of $[0, 1]$.

Hence, it is found that the exact membership $V_\mu(x)$ is bound between $V_t(x) \leq V_\mu(x) \leq 1 - V_f(x)$.

III. RELATED WORK

The earliest of work that was done in association rules and vague set was introduced by An Lu et al. They first proposed that every database contains some amount of vagueness. It is important to identify those objects that are vague on which vague set theory can be applied. For that purpose a concept of vague functional dependencies (VFDs) [7] was introduced which finds vagueness by using a measure called Similar Equality (SE_Q).

In 2007, the authors proposed a method to form vague relations in the database by using the well known integrity constraint Functional Dependency [8]. The best approximation of the vague relation is determined with respect to

functional dependencies by using the median and imprecision membership and by notion of *mi*-overlap between vague sets relations.

In another work, the author proposed a method to mine vague association rule (VARs) [9]. Herein the author finds vague rules by giving a notion of intent, attractiveness and hesitation. These parameters are derived from the median and imprecision memberships. Afterwards, the attractiveness (*A*) and hesitation (*H*) are used to form *AH*-pair database which is mined and rules are generated.

An Lu et al. also discussed about the hesitation status of any item being sold. The hesitation statuses (*HS*) [10] which are the sub parts of overall hesitation convey a lot of information. For example, an item whose hesitation is zero has the maximum attractiveness. Thus mining hesitation information is an important element to find VARs. Furthermore, the author also gives four types of support and confidence values. These values are based on the individual attractiveness (*A*), hesitation (*H*) and the *AH*- and *HA*-pairs database [9, 10].

Another work that was added in vague sets and association rules was done by Anjana Pandey et al. [13] where the domain specific vagueness was identified in the area of academics. The vague objects identified are courses and their relation with other vague object i.e. students and how they correlate to each other.

IV. PROPOSED WORK

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 methods that implements the core concept of true and false membership for finding support and confidence to generate vague rules.

Since, the vague set theory states that the precision of an object that either belongs to certain state can be defined by its true membership (V_t) and false membership (V_f). We propose that for a vague item there could be two possible states. Our notion of vagueness in database is based on the FMCG data collected. We consider that the products having multiple product codes due to different packaging leads to vagueness. Thus, if a product item is vague then it is bound to have more than one product item code. A vague item that is present in the transaction *T* is considered as its true membership (V_t), whereas the variant of that particular vague item that is found in the transaction, excluding the ones those are already found by its main variant item is considered as false membership (V_f). Thus when applying the vague set theory to association rules, the support boundaries lie within the interval $[0, 1]$ i.e. the minimum support boundary is defined as the total occurrence of a vague item in the database, its true membership (V_t), and the maximum support boundary is the negation of evidence of an item ($1 - V_f$). Thus this limits the support as $[(V_t), (1 - V_f)]$.

Algorithm: Vague Itemset minEr (VIE)

1. Vague Table Generation Algorithm

VagueTab (D)

- 1) Scan the database *D* to find the total number of distinct items;
- 2) For $i = 0, 1, 2, \dots$ 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 $j = 0, 1, 2, \dots$ 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*);

2. Vague Itemset minEr (VIE) Algorithm

VIE (D, V_t)

- 1) Calculate 1st vague frequent itemset by scanning *D*;
- 2) For $i=0, 1, 2, \dots t_n$ 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 2nd 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*);

3. 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);

V. EXPERIMENTS & RESULT

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 1		
		(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 2		
		(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 overmining and forms *puissant rules*. Thus vague sets when incorporated with association rules provide a contrasting meaning to the classical approach and gives better results.

VI. DISCUSSION AND FUTURE SCOPE

As it is perceptible from the above approach there is an important characterization that vague set theory could make in data mining. We have induced it with association rules and have found it to be acceptable in terms of the quality of rules generated. There is a huge possibility for extending this work in field of data mining. Since association rules are considered to be based on the market-basket analysis model, our approach finds importance in an area we call as Profit Pattern Mining. To include vagueness as part of rule generation mechanism gives better understanding for which market strategy should be considered as employable. Mining profit patterns either by value based or percentage based will find helpful in making decisions according to the market scenario.

VII. CONCLUSION

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.

VIII. REFERENCES

- [1] Agrawal Rakesh, Imielinski Tomasz & Swami Arun, "Mining Association Rules between Sets of Items in Large Databases" Proceeding SIGMOD '93 Proceedings of the 1993 ACM SIGMOD international conference on Management of data Pages 207-216 ACM New York, NY, USA ©1993
- [2] Agrawal Rakesh, Srikant Ramakrishnan, "Fast Algorithms for Mining Association Rules - A Priori", Proceeding VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases Pages 487-499 Morgan Kaufmann Publishers Inc. San Francisco, CA, USA ©1994
- [3] Arvind Terrence S., Badhe Vivek, "Comparative Analysis of Fuzzy, Rough, Vague & Soft set Theories in Association Rule Mining" 2014 © International Journal Of Scientific Progress And Research (IJSPR) ISSN: 2349 - 4689 Volume-02, Number- 01, Aug 2014
- [4] Arvind Terrence S., Badhe Vivek, "The Application of Vague Set Theory in Association Rule Mining: A Survey" 2015 © International Journal of Advanced Research in Computer Science (IJARCS) ISSN: 0976-5697 Volume 6, No. 1, Jan-Feb 2015
- [5] Badhe Vivek, Thakur R.S., Thakur G.S., " Vague Association Rule Mining for Profit Patterns and Decision Making", International Conference on Intelligent, Computational and Informative Systems, Goa(India), October'2014.PP76-78. ISSNs: 2249-071X.
- [6] Gau Wen-Lung and Buehrer Daniel J., "Vague Sets", © 1993 IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, VOL. 23, NO. 2, MARCH/APRIL 1993
- [7] Lu An, Ng Wilfred, "Managing Merged Data by Vague Functional Dependencies" 2004 © Springer.
- [8] Lu An, Ng Wilfred, "Handling Inconsistency of Vague Relations with Functional Dependencies" Conceptual Modeling - ER 2007 Lecture Notes in Computer Science Volume 4801, 2007, pp 229-244 © Springer
- [9] Lu An, Wilfred Ng, Yiping Ke, and James Cheng, "Mining Vague Association Rules" DASFAA 2007, LNCS 4443, pp. 891–897, 2007 © Springer-Verlag Berlin Heidelberg 2007
- [10] Lu An, Wilfred Ng, Yiping Ke, and James Cheng, "Mining Hesitation Information by Vague Association Rules" ER 2007, LNCS 4801, pp. 39–55, 2007 © Springer-Verlag Berlin Heidelberg 2007
- [11] Molodtsov D, "Soft Set Theory – First Results", Computers and Mathematics with Applications 37 (1999) 19-31 © Elsevier Inc.
- [12] Motro Amihai, "Management of Uncertainty in Database Systems", Pages 457-476 ACM Press/Addison-Wesley Publishing Co. New York, NY, USA ©1995
- [13] 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
- [14] Pawlak Zdzislaw, "Rough sets" International Journal of Computer & Information Sciences October 1982, Volume 11, Issue 5, pp 341-356 © 1982 Plenum Publishing Corporation.
- [15] Wierman Mark J., "An Introduction to the Mathematics of Uncertainty", August 2010, Center for the Mathematics of Uncertainty, Creighton University
- [16] Zadeh L., "Fuzzy Sets", INFORMATION AND CONTROL 12, 94-102 (1968).