ISSN No. 0976-5697

Volume 6, No. 1, Jan-Feb 2015

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

The Application of Vague Set Theory in Association Rule Mining: A Survey

Terrence Shebuel Arvind
M. Tech Scholar, Department of Computer Science
Engineering (CSE)
Gyan Ganga College of Technology
Jabalpur (M.P.), India

Vivek Badhe
Professor, Department of Computer Science
Engineering (CSE)
Gyan Ganga College of Technology
Jabalpur (M.P.) India

Abstract- Data Mining is one of the major research areas today. There are a number of techniques that were developed when it first emerged and with time the development of new techniques and enhancements over old ones are still in effect. One such area is Association Rules (ARs). Being the simple method of all there is a large possibility of advancing this method. There are a number of applications that have been developed for Association Rules; but considering the data integration from various sources and its heterogeneity some amount of vagueness persists. To deal with vague data existing in databases some proper technique must be weighed in. One such technique used, and the focus of our survey is Vague Set Theory and its application in Association Rule Mining.

Keywords: Data Mining, Association Rule (ARs) Mining, Vague Set Theory, Vague Functional Dependencies (VFDs)

I. INTRODUCTION

Data Mining is primarily known for its convenient approach in finding hidden patterns. Ever since the information gathering and storage is ameliorated, experts (commonly known as knowledge expert in data mining community) have encountered several challenges for how a fruitful, interesting, novel, hidden knowledge can be discovered (or mined) from such huge, vast and vague data.

Considering the above locus of data mining, we have a few approaches that are quite effective in discovering such hidden patterns. We refer them as data mining techniques. Each technique has its own merits over the other. One such technique used is *Association Rule Mining* [1], [2]. These rules are very simple as they are based on statistical occurrence of the data items.

As discussed above, the data collection is not only huge but vague too! Thus, some measures need to be taken which directly or indirectly deal with vagueness. Earlier, fuzzy theory was approached to deal with approximation and imprecision but it limits up to certain extent [8], [12]. Other approaches include Rough Set Theory, Vague Sets, and Soft Sets [3], [4], [5]. Our discussion is based upon the combination (or embedded) of association rules with vague set theory [6].

II. DEFINITION

A. *Association Rule Mining:*

In early years of the 90's at IBM, Rakesh Agrawal et al. [1], [2] proposed statistical methods which later became the most researched topic in Data Mining. The technique was named Association Rule Mining. The significance of this method was based on pure statistical occurrence of the data attributes in a large database. The frequent patterns are discovered and then put into an algorithm which further generate rules and give a certain relationship among different data attributes of database.

An association rule is probabilistic relationship, of the form of $A \Rightarrow B$ which are the sets of database attributes, where A and B are sets of individual items, and $A \cap B = \phi$. If a set of transactions *T* is given, then rules can be generated pertaining to certain constraints. These constraints (or measures) are called *support* and *confidence*. The *support* of the rule is the occurrence of any data attribute of set A (or B) to the total number of attributes a transactions *T* in a database *D* that satisfy the union of items in A and B. The probability, taken as the ratio of the transactions containing A also containing B, is termed as *confidence* of the rule.

Support and confidence helps to compute the association among different data attributes. Frequent patterns are the basis for any data mining task for which rules must be generated. These rules are application-specific and may vary according to a certain threshold; i.e. how much the strength of a rule should be considered.

There are many algorithms that have been developed and implemented that use association rule mining methodology. The foremost was named *AIS Algorithm* [1] which first introduced method for mining very large databases. Later, Rakesh Agrawal et al. [2] proposed a method that became a milestone in data mining research. The algorithm is called *A Priori* which uses support and confidence as its core components and creates large itemsets and rules where formed.

B. *Vague Set Theory:*

Since the introduction of fuzzy theory [3] there has been a lot of research in the direction of reasoning based computing. In following decades few other set theories [8], [12] developed all of which had the capability to deal with imprecision, approximation, uncertainty, and even vagueness.

Vague set theory [6] however, is similar to its predecessors but it bounds the intervals between 0 and 1. Unlike fuzzy, which provide a graded membership, vague set encloses that membership within interval (0, 1). Gau and Buehrer proposed that if V is a vague set in U which is the universe of discourse, then membership formed is within the interval *[0, 1]* as:

$$V_t\colon U \rightarrow [0, 1],$$

V_f: $U \rightarrow [0, 1]$, and
$V_t(x) + V_f(x) \leq 1$,

Where, V_t and V_f are true and false membership functions.

Also, $V_t(x)$ is a *lower bound* function on the grade of membership of U derived from the evidence for *x*, and $V_f(x)$ is a *lower bound* function on negation of *x* derived from the evidence against *x*.

Vague set is now being used in many research areas. Its importance was seen to deal with data which cannot be kept either in the true or false category. Such data is present in real world problems which could be computed by application of vague sets. But to identify vagueness is one challenging task because you need to differentiate the particular domain where it may be applied. Not all data collected may be vague. Fuzzy and other related set theories, all which have some commonness in them can contribute in resolving any problem but still they also differ in certain aspect [12].

An Lu et.al also discussed about how to apply certain theory to specific problem and which could be beneficial than other [8].

III. RELATED WORK

Fuzzy theory is primarily used in finding association rule. Fuzzy has the capability to deal with uncertainty and imprecision which is quite practicable in this area of discovering rule. The other theories [12] also have something to add to this research field but we limit our discussion on the application of vague set theory for generating more interesting rules. In our survey we find that a lot of work continues in the direction of other set theory (fuzzy, rough, soft sets) [3], [4], [5] but vague set theory is the least touched of all.

As discussed above, the two techniques are used together to increase the potency of the rules being generated. The foremost application of vague sets was introduced by An Lu et al. where vague association rules were generated [9]. The method comprises the task of finding vague information from the database which is under consideration. For that a certain additional measures were taken apart from support and confidence. The challenging work in such field of study is to identify vagueness. Not all information in databases could be vague, but with certain parameters it can be identified.

The additive parameters considered are meant to enhance the discovery of rules. For that reason, we decompose first how to determine vagueness in a database and for that Vague Functional Dependencies (VFDs) [7] is taken into account. Secondly, we discuss how this vague data can be determined for specific application databases by forming AH-pairs [9], [10] of the database transactions. And lastly, we discuss how one can represent the rules by defining additional types of support and confidence.

Firstly, to deal with vagueness one must first need to identify it. This identification of vague parameters can be seen in An Lu and Wilfred Ng work [7]. They proposed a mechanism to deal with data attributes of a database in a similar approach as is done by normal crisp set approach to databases, i.e., by managing functional dependencies. The only diverse notion is to address the vague functional dependencies. The first important factor that is covered is to find a similarity measure between any two vague values. This is done by finding the difference of difference between true and false membership and also by difference of sum between true and false membership of any two vague values.

$D_d = \left|\left(V^{x}_{t} - V^{y}_{t}\right) - \left(V^{x}_{f} - V^{y}_{f}\right)\right| / 2$ and

$S_d = \left|\left(V^{x}_{t} - V^{y}_{t}\right) + \left(V^{x}_{f} - V^{y}_{f}\right)\right|$

Where considering *x* and *y* to be vague values such that, $x = \left[V^{x}_{t}, \left(1 - V^{x}_{f}\right)\right]$ *and* $y = \left[V^{y}_{t}, \left(1 - V^{y}_{f}\right)\right]$ to be the subintervals. Now the similarity measure between these two vague values is given by:

$$SM(x,y) = \sqrt{(1 - D_d)(1 - S_d)} = \sqrt{\left(1 - \frac{\left|\left(V^{x}_{t} - V^{y}_{t}\right) - \left(V^{x}_{f} - V^{y}_{f}\right)\right|}{2}\right) . \left(1 - \left|\left(V^{x}_{t} - V^{y}_{t}\right) + \left(V^{x}_{f} - V^{y}_{f}\right)\right|\right)}$$

In addition, a distance measure between vague values *x* and *y* is defined by *D(x,y) = 1–SM(x,y)*.

Once the vagueness is identified in the database then similar equality (S_{EQ}) for vague relations r over scheme R (r ⊆R) for any two tuples tp and tq is given as:

$$S_{EQ}(t_p[A_i], (t_q[A_i]) = \frac{1}{n}\sum_{k=1}^{n} \sqrt{\left(1 - \frac{\left|\left(V_t\left(t_p[Ai](u_k)\right) - V_t\left(t_q[Ai](u_k)\right)\right) - \left(V_f\left(t_p[Ai](u_k)\right) - V_f\left(t_q[Ai](u_k)\right)\right)\right|}{2}\right) . \left(1 - \left|\left(V_t\left(t_p[Ai](u_k)\right) - V_t\left(t_q[Ai](u_k)\right)\right) + \left(V_f\left(t_p[Ai](u_k)\right) - V_f\left(t_q[Ai](u_k)\right)\right)\right|\right)}$$

Thus the Similar Equality of two tuples t_p and t_q on data attributes $X = \{A_1, A_{2,} \dots, A_n\}$ (X ⊆ R) in vague relation r is given by:

$$S_{EQ}(t_p[X], (t_q[X]) = S_{EQ}(t_p[A_1, \dots, A_n], (t_q[A_1, \dots, A_n]) = \min\{S_{EQ}(t_p[A_1], (t_q[A_1]), \dots, S_{EQ}(t_p[A_n], (t_q[A_n])\}$$

Thus, by above equations we can find the vagueness in database and can calculate the Vague Functional Dependencies (VFDs) in r over a relation schema $R = \{A_1, A_{2,} \dots, A_m\}$, where domain of $A_i, i = 1, \dots, m$ are vague sets. Once the VFDs are acknowledged, it is easier to facilitate to relate vague attributes to data mining task in forming association rules.

As VFDs are in the findings it is important to have contrasting measure between them because the values are still vague. For that, again An Lu et.al [10] proposed some membership functions that can be assigned as vague membership of a vague set. These are donated by median membership and imprecision membership. The median membership is defined as $M_m = \frac{1}{2}\left(V_t + \left(1 - V_f\right)\right)$ representing the overall evidence as $0 \leq M_m \leq 1$; whereas the imprecision membership is defined as $M_i = \left(\left(1 - V_f\right) - V_t\right)$ representing the overall imprecision if a vague value as $0 \leq M_i \leq 1$.

The author [10] also imparts additional information to these vague values to identify them more easily referred as Hesitation Status (HSs, being an online shopping scenario) of either ‘buying’ or ‘not buying’. With respect to HSs, we can obtain the intent of each item for some transaction which is also a vague value and lies within the subinterval $[V_t(x), (1 - V_f(\mathrm{x}))]$. The intent of an attribute is used to further define the attractiveness of an item representing the overall evidence and hesitation pairs of an item (or attribute) which has association with such other attributes of the database [9], [10].

The purpose behind inducing these AH-pair is to categories the database attributes as per vague set parameters. For attributes that follow (or has support) are categories as true membership (V_t) and the ones that are not (or is against) are called false membership (V_f), but this information does not tell us anything about the vagueness. The other category, i.e. the intent [12] of any attribute, neither true nor false, are calculated by two arguments as discussed above. The attractiveness (A) of any attribute t is calculated as $A(t) = \frac{1}{2}\left(V_t + \left(1 - V_f\right)\right)$ and hesitation (H) of any attribute t is $H(t) = \left(\left(1 - V_f\right) - V_t\right)$. The values from these attractiveness and hesitation pair are reduced as AH-pair transaction database [10].

Now, the vague association rules (VARs) [9] are generated by this AH –pair database. The rules that were generated were based on datasets of an online shopping scenario such as Amazon.com where the “sold”, “not sold” and “almost sold” status events were taken with respect to the support, against and hesitation information of an item. Thus if we have interestingness and hesitation of an item, we can analyze it to improve the latter information. Hence, AH –pair support and confidence are derived. The four types of AH–support and confidence to evaluate VARs are:

Support for an AH–pair database D, where $X \cup Y = Z$, the number of itemsets are as follows:

a. The A–support of Z, denoted and defined as *Asup(Z)* = $\frac{\sum_{T\in D}\prod_{z\in Z} M_A(z)}{|D|}$
b. The H–support of Z, denoted and defined as *Hsup(Z)* = $\frac{\sum_{T\in D}\prod_{z\in Z} M_H(z)}{|D|}$
c. The AH–support of Z, denoted and defined as *AHsup(Z)* = $\frac{\sum_{T\in D}\prod_{x\in X, y\in Y} M_A(x)M_H(y)}{|D|}$
d. The HA–support of Z, denoted and defined as *HAsup(Z)* = $\frac{\sum_{T\in D}\prod_{x\in X, y\in Y} M_H(x)M_A(y)}{|D|}$

Similarly, confidence for a VAR, $r = (X \Longrightarrow Y)$ in AH−pair database D, where $X \cup Y = Z$

a. The A–confidence of *r* when both X and Y are *A FIs* is, *Aconf(r)* = $\frac{Asup(Z)}{Asup(X)}$
b. The H–confidence of *r* when both X and Y are *H FIs* is, *Hconf(r)* = $\frac{Hsup(Z)}{Hsup(X)}$
c. If X is an A FI and Y is an H FI, then AH–confidence of r is, *AHconf(r)* = $\frac{AHsup(Z)}{Asup(X)}$
d. If X is an H FI and Y is an A FI, then HA–confidence of *r* is, *HAconf(r)* = $\frac{HAsup(Z)}{Hsup(X)}$

By implying four types of support and confidence in our study, more valid VARs can be deduced.

Another application [11] of vague sets used in association rule was discussed by Anjana Pandey and K. R. Pardasani. The objective was to discover the vagueness in area of academia. As discussed above, to identify vagueness is not an easy task. Thus, the application deals with the association between students and the courses they intend to attend. If any number of students (S_i) takes a certain course (C_i), then vagueness can be identified by determining the students who regularly attend the course classes (support) and those who not (against). The students who hesitate to attend a course could be determined by the intent value of student’s attractiveness and hesitation, forming the AH –pair transaction database after which the algorithm used will discover the formidable rules.

IV. DISCUSSION AND FUTURE SCOPE

So far it is evident that in contemporary world as more and more information is rendered by computerized systems; there will always be a need to extract knowledge from huge data. But what if the data obtained is vague? In such case vague set theory could be used and dealt with. However, the survey enlightens us that the practical implementation is quite multifactorial. We ascertain by going over the examples and problems in our survey that the application might not be restrained on singleton variable (attribute) that partakes in the rule making. What if the vagueness could be extended for two attribute (or multi-valued)? In response to this, if we employ vagueness in context of multi-valued attributes, not only will it consent to yield more vague rules rather would also impart much knowledge in discovering patterns that may ultimately govern in decision making. But, to find single (or multiple) vague value in any database it is essential to have prior knowledge of the domain since the significance of vagueness differ as the domain varies. Thus we concur that it is impossible to conjure up that a particular method could, would, or should be pertinent in any application.

Henceforth, we intend and justify that vagueness should be dealt with specifics, meeting a certain criteria and not being generic in nature. For sake of argument, in FMCG data the vagueness superficially identified are Quantity, Price, Temporal, Product name/code, Profit, to name a few. While in medical data such as Pathology, the vagueness might be Symptoms, Diseases and Genetic Disorders. Besides these more could be identified later when correlating to other attributes which are still obscure.

Another aspect our survey has inclination towards is the measures used in rule generation: support and confidence. Since we propose a notion of determining vagueness in multi-valued attributes because of which the generic measures may fail or neglect the interestingness of a rule. If reasoned carefully we find that for every vague value provide different support values which can addresses vagueness in much astute manner.

V. CONCLUSION

The above survey concludes that vague sets are an important mathematical tool that has an extensive use in discovering vagueness in any data, if applicable, which could be found in most of today’s information disseminating systems that are automated and generate large amount of data. If used with association rule mining whose primary goal is to discover rules that would help increase the original concept of the market-basket analysis and somewhat in decision making process. The rules that are formed are vague in nature but do allow us a new perspective in association rule mining.

VI. REFERENCES

[1] Rakesh Agrawal, Tomasz Imielinski & Arun Swami, “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] Rakesh Agrawal, Ramakrishnan Srikant, “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] L. Zadeh, “Fuzzy Sets”, INFORMATION AND CONTROL 12, 94-102 (1968).

[4] Zdzislaw Pawlak, “Rough sets” International Journal of Computer & Information Sciences October 1982, Volume 11, Issue 5, pp 341-356 © 1982 Plenum Publishing Corporation.

[5] D Molodtsov, “Soft Set Theory – First Results”, Computers and Mathematics with Applications 37 (1999) 19-31 © Elsevier Inc.

[6] Wen-Lung Gau and Daniel J. Buehrer, “Vague Sets”, © 1993 IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, VOL. 23, NO. 2, MARCHIAPRIL 1993

[7] An Lu, Wilfred Ng, “Handling Inconsistency of Vague Relations with Functional Dependencies” Conceptual Modeling - ER 2007 Lecture Notes in Computer Science Volume 4801, 2007, pp 229-244 © Springer

[8] An Lu and Wilfred Ng, “Vague Sets or Intuitionistic Fuzzy Sets for Handling Vague Data- Which One Is Better?” ER 2005, LNCS 3716, pp. 401–416, 2005 © Springer-Verlag Berlin Heidelberg 2005

[9] An Lu, Yiping Ke, James Cheng, and Wilfred Ng, “Mining Vague Association Rules” DASFAA 2007, LNCS 4443, pp. 891–897, 2007 © Springer-Verlag Berlin Heidelberg 2007

[10] An Lu, Yiping Ke, James Cheng, and Wilfred Ng, “Mining Hesitation Information by Vague Association Rules” ER 2007, LNCS 4801, pp. 39–55, 2007 © Springer-Verlag Berlin Heidelberg 2007

[11] 2012 International Journal of Computer Applications (0975 – 8887) Volume 58– No.20, November “A Model for Mining Course Information using Vague Association Rule” Anjana Pandey, UIT RGPV Bhopal, K.R.Pardasani MANIT Bhopal

[12] Terrence S. Arvind, Vivek Badhe, “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