TOWARDS A LINGUISTIC APPROACH TO ASSESS FUZZY ASSOCIATION RULES

Miguel Delgado Nicolás Marín Daniel Sánchez[1] María-Amparo Vila
Department of Computer Science and Artificial Intelligence
University of Granada, E.T.S.I.I., Avda. Andalucia 38, 18071 Granada, Spain
e-mail: mdelgado@ugr.es, {nicm, daniel, vila}@decsai.ugr.es

Extended Abstract

Nowadays, databases are not only considered as stores for data, but also as a source of knowledge that can help organizations to improve their management and to reach their goals. This fact has motivated an increasing research in the field of Knowledge Discovery during the last ten years. Knowledge discovery is concerned with obtaining novel, previously unknown and potentially useful knowledge from databases. Many results in this area fall into data mining, the task of the knowledge discovery process that attempts to find novel and significant patterns in data.

One factor that determines the kind of patterns we can mine is the structure of data. In many cases, specially in the business environment, records in data are perceived as transactions, a transaction being a subset of a (finite) set of items I. Databases are then perceived as sets of transactions, we call T-sets. A classical example are market basket data. Each basket is a transaction that contains a subset of products (items). The set of all the baskets bought by clients is a T-set.

Typical patterns in T-sets are association rules [1]. Association rules are "implications" that relate the presence of items in transactions. The classical example are rules extracted from market baskets. Association rules relate the presence of items in the same basket, for example "every basket that contains milk contains bread", usually noted *milk* $\Rightarrow$ *bread*.

When mining for association rules it is important to measure both its interest and its accuracy. The usual measures to assess association rules are support and confidence, both based on the concept of support of an *itemset* (a subset of items). Given a set of items I and a T-set R on I, the support of an itemset $I_0 \subseteq I$ is

$$supp(I_0) = \frac{|\{\tau \in R \mid I_0 \subseteq \tau\}|}{|R|}$$

The support of a rule $A \Rightarrow C$ is

$$Supp(A \Rightarrow C) = supp(A \cup C)$$

and its confidence is

$$Conf(A \Rightarrow C) = \frac{supp(A \cup C)}{supp(A)} = \frac{Supp(A \Rightarrow C)}{supp(A)}$$

Support is the percentage of transactions where the rule holds. Confidence is the conditional probability of C with respect to A or, in other words, the relative cardinality of C with respect to A.

Our work involves a natural extension of the concept of transaction; we call it *fuzzy transaction*, and it is defined as a fuzzy subset of a set of items. On this basis we define an FT-set as a (crisp) set of fuzzy transactions. These concepts lead to the definition of *fuzzy association rule* as an association rule that hold in an FT-set. Fuzzy association rules are very useful in solving several interesting applications [3], such as finding association rules between quantitative attributes in relational databases, among others.

One of the main problems with respect to fuzzy association rules is how to assess them, i.e., how to measure their interest and accuracy, because the set of transactions that contain a given item

[1]Corresponding author.

is fuzzy. Hence, a solution based on support and confidence has to deal with cardinality of fuzzy sets. In previous works we have studied a numerical solution to this problem [3, 4]. In this paper we present a methodology to assess fuzzy association rules by means of linguistic terms, based on some of the tools we developed for a numerical approach. We consider the linguistic approach as an step towards providing the user with more understandable knowledge with richer semantics.

Fuzzy Transactions and Fuzzy Rules

The following is a formalization of the concepts sketched before, that can be found in [3]. Given a (finite) set of items I, we call fuzzy transaction to any fuzzy subset $\tilde{\tau} \subseteq I$, $\tilde{\tau} \neq \emptyset$. For every $i \in I$ we note $\tilde{\tau}(i)$ the membership degree of i in $\tilde{\tau}$. We use the same notation for the membership degree of an itemset $I_0 \subseteq I$ to a fuzzy transaction $\tilde{\tau}$, $\tilde{\tau}(I_0)$. We define

$$\tilde{\tau}(I_0) = \min_{i \in I_0} \tilde{\tau}(i)$$

Given an FT-set T based on I, we call *representation* of an itemset I_0 to the fuzzy set $\tilde{\Gamma}_{I_0} \subseteq T$, defined as

$$\tilde{\Gamma}_{I_0} = \sum_{\tilde{\tau} \in T} \tilde{\tau}(I_0)/\tilde{\tau}$$

A fuzzy association rule is a link of the form $A \Rightarrow C$ such that $A, C \subset I$ and $A \cap C = \emptyset$. The itemsets A and C are called *antecedent* and *consequent* respectively. The rule $A \Rightarrow C$ holds with total accuracy in T when $\tilde{\Gamma}_A \subseteq \tilde{\Gamma}_C$. We are interested in finding rules whose support and confidence are greater than two user-defined thresholds *minsupp* and *minconf*.

A numerical approach to assess fuzzy association rules

We proposed in [4] a numerical approach based on the evaluation of quantified sentences of the form "Q of F are G" [10], where Q is a linguistic quantifier, and F and G are fuzzy subsets of a finite reference set X. We focus on relative quantifiers, linguistic terms that represent fuzzy percentages, that are defined by means of fuzzy sets on $[0, 1]$. Examples are "most", "almost all" or "many".

We define the support of an itemset I_0 in an FT-set T as the evaluation of the quantified sentence

$$Q_M \text{ of } T \text{ are } \tilde{\Gamma}_{I_0}$$

while the support of a rule $A \Rightarrow C$ in T is the evaluation of

$$Q_M \text{ of } T \text{ are } \tilde{\Gamma}_A \cap \tilde{\Gamma}_C$$

and its confidence is the evaluation of

$$Q_M \text{ of } \tilde{\Gamma}_A \text{ are } \tilde{\Gamma}_C$$

where Q_M is a quantifier defined as $Q_M(x) = x$ $\forall x \in [0, 1]$.

The evaluation of a quantified sentence yields an accomplishment degree, usually a compatibility degree between the relative cardinality of G with respect to F and the quantifier. We employ the method GD [5]. The evaluation of "Q of F are G" by means of GD, $GD_Q(G/F)$, yields

$$\sum_{\alpha_i \in \Delta(G/F)} (\alpha_i - \alpha_{i+1})\, Q\left(\frac{|(G \cap F)_{\alpha_i}|}{|F_{\alpha_i}|}\right) \qquad (1)$$

where $\Delta(G/F) = \Lambda(G \cap F) \cup \Lambda(F)$, $\Lambda(F)$ being the level set of F, and $\Delta(G/F) = \{\alpha_1, \ldots, \alpha_p\}$ with $\alpha_i > \alpha_{i+1}$ for every $i \in \{1, \ldots, p\}$. The set F is assumed to be normalized. If not, F is normalized and the normalization factor is applied to $G \cap F$.

The quantifier Q_M is unique because it has the feature that, when T is a T-set, the evaluation of the sentences are the usual measures of support and confidence of crisp association rules, provided that the evaluation method verifies the intuitive property that, when A and C are crisp, the evaluation of "Q of F are G" is

$$Q\left(\frac{|F \cap G|}{|F|}\right)$$

(method GD verifies this property [5]). Hence, we can interpret the usual measures of confidence and support as the degree to which the confidence and support of an association rule is Q_M. Other properties of the generalization of the support/confidence framework by means of method GD with the quantifier Q_M are studied in [3].

Also in [3] we provide an algorithm to perform evaluations by means of GD in time $O(|T|)$, and an adaptation of a basic algorithm to find fuzzy association rules. They have been successfully employed to mine fuzzy association rules in large relational databases [4].

A linguistic approach

The approach we propose here relies in a basic set P of linguistic terms, specifically relative quantifiers. There are some widely accepted points we assume. Humans are not used to deal with many values in a domain, so we shall employ a small number of quantifiers. Also it is usual to employ an even number of terms, and to define P such that for every $Q \in P$, the antonym of Q

$$antQ(x) = Q(1-x) \quad \forall x \in [0,1]$$

is also in P. Another reasonable property is that every value $x \in [0,1]$ can belong in some degree to at most two quantifiers. Moreover, P must be a fuzzy partition of $[0,1]$ in the sense that

$$\sum_{Q \in P} Q(x) = 1 \quad \forall x \in [0,1]$$

Finally, we shall use symmetric triangular-shaped quantifiers. We call Q-set a set of quantifiers verifying the aforementioned properties. Figure 1 shows a simple Q-set we call P_5.

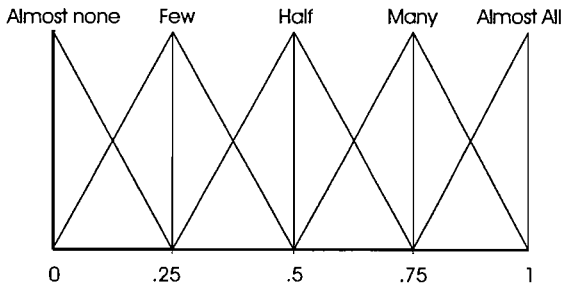

Figure 1: The Q-set P_5

We define support and confidence as second order fuzzy sets on a given Q-set P as follows:

Definition 1 *The support of $I_0 \subseteq I$ on T is*

$$supp^P(I_0) = \sum_{Q \in P} GD_Q\left(\left(\widetilde{\Gamma}_{I_0}\right)/T\right)\Big/ Q \qquad (2)$$

Definition 2 *The support of $A \Rightarrow C$ on T is* $Supp^P(A \Rightarrow C) = supp^P(A \cup C)$.

Definition 3 *The confidence of $A \Rightarrow C$ on T is*

$$Conf^P(A \Rightarrow C) = \sum_{Q \in P} GD_Q\left(\widetilde{\Gamma}_C/\widetilde{\Gamma}_A\right)\Big/ Q \quad (3)$$

This kind of representation of fuzzy information is highly accepted and very extended, see [6, 8].

Example 1 *This example is from [3]. Table 1 shows the FT-set T_6 on $I = \{i_1, i_2, i_3, i_4\}$:*

Table 1: The set T_6 of fuzzy transactions

	i_1	i_2	i_3	i_4
$\tilde{\tau}_1$	0	0.6	0.7	0.9
$\tilde{\tau}_2$	0	1	0	1
$\tilde{\tau}_3$	1	0.5	0.75	1
$\tilde{\tau}_4$	1	0	0.1	1
$\tilde{\tau}_5$	0.5	1	0	1
$\tilde{\tau}_6$	1	0	0.75	1

Rows are fuzzy transactions, for example $\tilde{\tau}_1 = 0.6/i_2 + 0.7/i_3 + 0.9/i_4$. Columns can be seen as representations of items, for example $\widetilde{\Gamma}_{i_1} = 1/\tilde{\tau}_3 + 1/\tilde{\tau}_4 + 0.5/\tilde{\tau}_5 + 1/\tilde{\tau}_6$.

Let's assess the rule $\{i_1, i_2\} \Rightarrow \{i_4\}$ by using the Q-set P_5 of figure 1. The linguistic approach yields:

- $Supp^{P_5}(\{i_1, i_2\} \Rightarrow \{i_4\}) = \frac{1}{2}/Almost\ None + \frac{1}{3}/Few + \frac{1}{6}/Half$
- $Conf^{P_5}(\{i_1, i_2\} \Rightarrow \{i_4\}) = 1/AlmostAll$

The thresholds minsupp and minconf are going to be based on the same linguistic representation that support and confidence. However, we allow the user to provide a number if preferred, from which we can obtain the representation by evaluating the quantifiers. For example, from the numerical threshold 1/3 we obtain $\frac{2}{3}/Few + \frac{1}{3}/Half$.

Given a representation $\sum_{Q_i \in P} \alpha_i/Q_i$ with P a Q-set, we define its center as the number $\sum_{Q_i \in P} \alpha_i k_i$, where k_i is the kernel of Q_i (and hence a number, since Q_i are triangular-shaped). In general, given a Q-set P, a value $x \in [0,1]$ is equivalent to a representation $Q_1(x)/Q_1 +$

$Q_2(x)/Q_2$ ($Q_1, Q_2 \in P$ being consecutive quantifiers), since its center is x.

We employ the center to decide if support and confidence are greater than the thresholds by comparing the center of the measure and the center of the threshold. However, this procedure is transparent to the user, and it has the feature that it generalizes the procedure in the crisp case. Indeed, if T is crisp and the numerical support of $A \Rightarrow C$ is s, then $GD_Q(A \cup C/T) = s$ [5] and hence the linguistic representation will be $Q_1(s)/Q_1 + Q_2(s)/Q_2$, whose center is s. This holds also for confidence. We omit the proofs because of the lack of space.

The linguistic approach can be easily incorporated into the algorithm described in [3], since the only change is to evaluate a quantified sentence for every $Q \in P$ instead of doing it with Q_M only. Since $|P|$ is fixed and small, the complexity of the process keep being the same. As the experiments we performed in [4] showed that time and space spent using the aforementioned algorithm were both acceptable, we expect a similar result in the experiments with the linguistic approach.

Related Work

Fuzzy association rules as patterns discovered in data have been studied by other authors, though in the specific setting of relational databases only, see [2, 4, 7] among others. A brief review of the publications about this topic can be found in [3]. However, all the existing approaches to assess fuzzy association rules are based on numerical quantities. Another related work is [9], where the authors introduce a methodology to discover what they call *quantified fuzzy rules*; these are rules assessed by means of a single quantifier.

Conclusions

We have presented a methodology to assess fuzzy association rules by using linguistic terms. We think that, this way, the assessment of rules is closer to that of humans. We have focused on confidence and support only, but we shall work towards an approach based on better measures, such as the certainty factors we employed in [3, 4].

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. Of the 1993 ACM SIGMOD Conference*, pages 207–216, 1993.

[2] W.H. Au and K.C.C. Chan. FARM: A data mining system for discovering fuzzy association rules. In *Proc. FUZZ-IEEE'99, Seoul, South Korea, Vol. 3, Pp. 22-25*, 1999.

[3] M. Delgado, N. Marín, D. Sánchez, and M.A. Vila. Fuzzy association rules: General model and applications. *IEEE Transactions on Fuzzy Systems*, 2001. Submitted.

[4] M. Delgado, D. Sánchez, and M.A. Vila. Acquisition of fuzzy association rules from medical data. In S. Barro and R. Marín, editors, *Fuzzy Logic in Medicine*. Physica-Verlag, 2000. To appear.

[5] M. Delgado, D. Sánchez, and M.A. Vila. Fuzzy cardinality based evaluation of quantified sentences. *International Journal of Approximate Reasoning*, 23:23–66, 2000.

[6] M. Delgado, J.L. Verdegay, and M.A. Vila. A linguistic version of the compositional rule of inference. In *SICICA'92*, pages 141–148, 1992.

[7] Chan-Man Kuok, Ada Fu, and Man Hon Wong. Mining fuzzy association rules in databases. *SIGMOD Record*, 27(1):41–46, 1998.

[8] W. Pedrycz. Applications of fuzzy relational equations for method of reasoning in presence of fuzzy data. *Fuzzy Sets and Systems*, 16:163–175, 1985.

[9] M. Umano, T. Okada, I. Hatono, and H. Tamura. Extraction of quantified fuzzy rules from numerical data. In *Proc. FUZZ-IEEE'2000*, pages 1062–67, 2000.

[10] L. A. Zadeh. A computational approach to fuzzy quantifiers in natural languages. *Computing and Mathematics with Applications*, 9(1):149–184, 1983.