AN EXTENSION OF ASSOCIATION RULES USING FUZZY SETS

Jee-Hyong Lee , Hyung Lee-Kwang

CS Dept. KAIST(Korea Advanced Institute of Science and Technology) Kusong, Yousong, Taejon, 305-701, Seoul Korea, Email:leejh@monami.kaist.ac.kr

Abstract. Association rules are a class of regularities existing between binary data tuples. This paper proposes an extension of association rules which can be applied to real-valued tuples. It discovers and describes association rules among real-valued tuples using fuzzy sets. The proposed method needs user-defined fuzzy sets for describing association rules. It extends the given tuples using the fuzzy sets and converts the extended tuples into binary tuples. Finally, it finds association rules by applying the existing algorithms for binary tuples to the converted binary tuples.

Keywords: Data mining, Association rules, Fuzzy sets

1 Introduction

Data mining is the technique which extracts the previously unknown and potentially useful information from large amount of data [1], [2]. Discovering association rules is one of the data mining techniques. Association rules give simple but strong knowledge on binary data tuples. They are the description that the tuples having a certain set of attributes also has another certain attribute. Association rules can be applied to various areas, such as analysis of customer's purchasing patterns.

However, most researches on association rules are devoted to binary data tuples. Most of data in the real world have real-valued attributes, so the binarization of those data may lose some information which the original has. For example, the data on customer's purchase usually have the information on the item and the amount. However, if we binarize the data by changing all amounts greater than 0 into 1, that is, keeping only the purchased items, we would lose the information on the amount of the items. The binarized data will have only the items which customers purchase.

This paper proposes the extended association rules which describe the association among real-valued attributes using fuzzy sets and the method discovering them. In the next section, we will describe briefly association rules and the discovery method, and then we will propose the extended association rule. Finally the experimental result and the concluding remarks are following.

2 Association Rules

In this section, we describe the association rules existing between binary data, and the algorithms for finding them [2], [3]. Let $R = \{I_1, I_2, \ldots, I_n\}$ be the set of attributes, and $t = \{t_1, \ldots, t_n\}, t_i \in \{0, 1\}$ be the tuple of the given schema $R = \{I_1, I_2, \ldots, I_n\}$. We will use $t[I_i]$ to denote the value of the attribute I_i in the tuple t, then $t[I_i] = 1$ represents that the value of the attribute I_i is 1, or the attribute I_i is in the tuple t. We also use $t[W] = \overline{1}$ where all $I \in W$ $(W \subseteq R)$ satisfy $t[I] = 1$. For $W \subseteq R$; $B \in R$, WB stands for $W \cup \{B\}$. We let $m(W)$ denote the set of tuples satisfying $t[W] = \overline{1}$. $|A|$ is defined as the number of items of set A.

Association rules are syntactically defined as $W \Rightarrow B$, $(W \subseteq R, B \in (R - W))$. The semantics is that the tuples satisfying $t[W] = \overline{1}$ also satisfies $t[B] = 1$. In other words, if the attribute set W is in a tuple, then the attribute B is also in the tuple. Because, in the real-world data, it seems to be very rare that all tuples satisfying $t[W] = \overline{1}$ also satisfies $t[B] = 1$, the conditions are usually relaxed as the following.

In case where n is the number of tuples, and two real number, γ, σ are given, if

$$|m(WB)| \geq \sigma n$$

and,

$$\frac{|m(WB)|}{|m(W)|} \geq \gamma$$

then, we say that the given tuples satisfy the association rule $W \Rightarrow B$. We call γ the confidence threshold and σ the support threshold. We consider an association rule is meaningful if the attributes of the association rule appear many times and there is a strong connection between the left-side attributes and the right-side attribute in the given tuples.

The meaning of association rules is simple and strong. For example, in the data on customer's purchase, let's suppose we have discovered the following association rule with $\gamma = 0.9$, and $\sigma = 0.3$:

Hamburger $\Rightarrow$ Coke.

Then, we can know that more than 30% of customers buy hamburger and coke together, and more than 90% of customers who buy hamburger also buy coke.

The algorithm for finding association rules consists of two phases.

1. Finding all subsets satisfying the given support threshold. Let's call those subsets covering sets.
2. Selecting the association rules which satisfy the confidence threshold among the association rules which are derived from the found covering sets.

Let's investigate it with an example. Let's suppose $R = \{A, B, C, D, E, F, G, H\}, \gamma = 0.9, \sigma = 0.3$, and the following tuples are given.

	A	B	C	D	E	F	G	H
t_1	1	1	1	1	0	0	0	0
t_2	1	1	1	0	1	0	0	0
t_3	1	1	0	0	0	1	0	0
t_4	0	1	1	0	0	0	1	0
t_5	1	1	0	0	0	0	0	1

A subset of R should appear in more than $1.5 (= 5 \times 0.3)$ tuples to be a covering set. Among the subset of R, $\{A\}$, $\{B\}$, $\{C\}$, $\{A, B\}$, $\{A, C\}$, $\{B, C\}$ and $\{A, B, C\}$ are covering sets. The association rules followed from each covering set are $A \Rightarrow C$, $C \Rightarrow A$ from $\{A, C\}$; $B \Rightarrow C$, $C \Rightarrow B$ from $\{B, C\}$; $AB \Rightarrow C$, $AC \Rightarrow B$, $BC \Rightarrow A$ from $\{A, B, C\}$. Among the candidate association rules, only three $A \Rightarrow B$, $C \Rightarrow B$ and $AC \Rightarrow B$ satisfy the confidence threshold. Therefore, the three association rules are accepted as meaningful among the given data tuples in case of $\gamma = 0.9$ and $\sigma = 0.3$.

The algorithm largely depends on the number of attributes and tuples. To find more efficient algorithm, some researches had been devoted to this problem and more faster algorithm was reported [5], [6].

3 Extended Association Rules

In this section, we propose association rules describing the associations between real-valued attributes, and the finding method. We define the association rules between real-valued attributes as the description of attribute-value co-occurrences, and we call it the extended association rules.

That is, the usual association rules are applied to binary data and describe the co-occurrence of attributes, but the extended association rules are applied to real-valued data and give information on the co-occurrence of the values of attributes. For example, in the customer's purchasing data, the extended association rules will give us something like the following rule :

(Hamburger, \$5) $\Rightarrow$ (Coke, \$2)

which says that the customers buying 5 dollars worth of hamburger have a tendency to buy 2 dollars worth of coke.

As we can notify from the example, the extended association rules deal with the attribute-value pairs, that is, it discovers the association rules between attribute-value pairs. Thus the questions related with the extended association rule are how to deal with the attribute-value pairs and how to discover the association rules among them. In binary data tuples, attributes take 0 or 1, so it is possible to list all attribute-value pairs and check them if they satisfy the given conditions. However, real-valued attributes will take one of infinitely many numbers, so that it is impossible to list all possible attribute-value pairs.

To cope with the problem, we use fuzzy set to describe real values and discover association rules, and we basically use the algorithms for the usual association rules. In the next section, we will present in detail how to describe the extended association rules with fuzzy sets and how to discover them.

3.1 Fuzzy sets of attributes

In order to deal with real values, we use fuzzy set. We assume that the fuzzy sets of each attribute are given by users. We use the fuzzy sets in describing and discovering association rules.

By using fuzzy sets, we reduce the number of attribute-value pairs. Fuzzy sets also make the description of the association rule concise and generalized. For example, if we have the following association rules,

(Hamburger, \$5) $\Rightarrow$ (Coke, \$2)
(Hamburger, \$6) $\Rightarrow$ (Coke, \$3)
(Hamburger, \$4) $\Rightarrow$ (Coke, \$1.5)

it can be written like this :

(Hamburger, Medium) $\Rightarrow$ (Coke, Small)

They also help users to easily understand the relations between attributes because the association rules can be presented in linguistic forms. Another merit of using fuzzy sets is that the user can control the generalization level or the detail level of association rules. If a user define many fuzzy sets for an attribute, then he will get the association rules which precisely describe the attribute with the fuzzy sets. If he reduces the number of fuzzy sets, he will get more generalized rules. A user can change the number of fuzzy sets according to his interest or the importance of an attribute, and it results in the different level of detail.

3.2 Extension and Binarization of tuples

As a pre-processing step of discovering the extended association rules, we extend the real-valued tuples into attribute-value pair tuples using fuzzy sets given by users. The extended association rules are based on attribute-value pairs.

If we have real-valued tuples and the schema R of them, the schema M of the extended tuples consists of $(I_i, f^j_{I_i})$, where $I_i \in R$ and $f^j_{I_i}$ is one of the defined fuzzy sets for I_i. The attribute-value pairs have the roles of attributes in M, and they represent the degree to which I_i is included in $f^j_{I_i}$. Thus $(I_i, f^j_{I_i})$ takes the value in $[0, 1]$.

For example, let's suppose the schema $R = \{I_1, I_2\}$ is given, and the fuzzy sets of I_1 are $f^1_{I_1}$, $f^2_{I_1}$, and $f^3_{I_1}$ and for I_2, $f^1_{I_2}$ and $f^2_{I_2}$ are defined. Then, the extended schema of R is

$$M = \{(I_1, f^1_{I_1}), (I_1, f^2_{I_1}), (I_1, f^3_{I_1}), (I_2, f^1_{I_2}), (I_2, f^2_{I_2})\}.$$

If we have a tuple of the schema R, $t = \{t_1, t_2\}$, the extended tuple is

$$m = \{\mu_{f^1_{I_1}}(t_1), \mu_{f^2_{I_1}}(t_1), \mu_{f^3_{I_1}}(t_1), \mu_{f^1_{I_2}}(t_2), \mu_{f^2_{I_2}}(t_2)\}.$$

$\mu_f(x)$ is the membership degree to which x is included in the fuzzy set f. Since an extended tuple consists of the membership degrees to which the values in the original tuple matches the fuzzy sets given by users, we call them the membership tuples.

Next, we get the binary tuples from the membership tuples. For discovering the extended association rules, we basically use the existing discovering method which manipulate binary data tuples. To apply the discovery method to real-valued tuples, we should convert the tuples into binary tuples. We convert the membership tuples to binary tuples using the membership degree threshold μ which is given by users. The values greater than μ in the membership tuples are converted into '1', otherwise into '0'. For example, $\mu = 0.3$ and a membership degree tuple, $\{0.4, 0.6, 0, 0.8, 0.2\}$, is given, then the binary membership tuple is $\{1, 1, 0, 1, 0\}$. We consider that a value satisfies a fuzzy set only when the value is included in the fuzzy set more highly than the membership degree threshold μ.

The next step is discovering association rules by applying the existing algorithm to the binarized tuples.

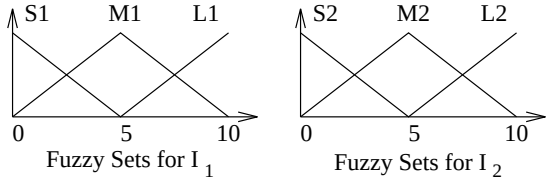
3.3 Discovery of association rules

Since we have converted the given real-valued tuple into binary, we can apply the existing discovering algorithm to them. If we apply the algorithm to the binary membership tuples, we easily get the association rules which describe the relations of attribute-value pairs. The discovered association rules can be classified into the two groups :

- The association rules between the fuzzy sets of different attributes (e.g.:$(I_1, f^1_{I_1}) \Rightarrow (I_2, f^1_{I_2})$)
- The association rules between the fuzzy sets of one attribute (e.g.:$(I_1, f^1_{I_1}) \Rightarrow (I_1, f^2_{I_1})$)

The first association rules say that a certain fuzzy set of an attribute usually appears when certain fuzzy sets of other attributes appear in a tuple. For example, "If the attribute A has large value, then the attribute B has medium value". However, the second describes somewhat useless fact, such as "If the attribute A has large value, then the attribute A has medium value". This is not an intended association rule and it gives little information to users. The reason why such association rules are generated is that we apply the existing algorithm to the binary membership tuples without any modification. The existing algorithms will discover the association rules between all attributes, but we do not need discover the association rules between attribute-value pairs of one attribute. That is, the existing algorithms regard the elements in tuples indicating different attributes, but in the binary membership tuples, some elements probably indicate the different aspects of an attribute. For example, in the previously defined membership schema M, $(I_1, f^1_{I_1})$, $(I_1, f^2_{I_1})$, $(I_1, f^3_{I_1})$ are for the attribute I_1. They are the different aspects which the value of I_1 can take, so we do not need to discover the association rules between them. Because such association rules are little use, we can eliminate them. The elimination also helps the discovering process by reducing the search space.

Let's investigate the whole process with an example. Let's suppose $R = \{I_1, I_2\}$ and {8,7}, {3,6} are given as the tuples of R. The fuzzy sets for I_1, I_2 are defined like the following figures.



The membership schema is

$$M = \{(I_1, S1), (I_1, M1), (I_1, L1), (I_2, S2), (I_2, M2), (I_2, L2)\}.$$

The membership tuples and the binary membership tuples with $\mu = 0.3$ are

$$(8, 7) \rightarrow \{0, 0.4, 0.6, 0, 0.6, 0.4\} \rightarrow \{0, 1, 1, 0, 1, 1\}$$

$$(3, 6) \rightarrow \{0.4, 0.6, 0, 0, 0.8, 0.2\} \rightarrow \{1, 1, 0, 0, 1, 0\}.$$

In case that $\sigma = 0.6$, and $\gamma = 0.9$, the covering sets are $\{(I_1, M1)\}, \{(I_2, M2)\}, \{(I_1, M1), (I_2, M2)\}$ and the association rules are

$$(I_1, M1) \Rightarrow (I_2, M2), \text{ and } (I_2, M2) \Rightarrow (I_1, M1).$$

Those say that "if the attribute I_1 is medium, then I_2 is also medium" and the reverse.

4 Experimental Results

In this section, we apply the proposed approach to real-valued data tuples and show the result. The used data tuples*) are on the housing data composed of 14 attributes. For the experiment, we choose only 11 attributes among them. They are the pupil-teacher ratio(PTRATIO), the proportion of non-retail business acres(INDUS), the weighted distances to the five employment centers(DIS), the median value of owner-occupied homes(MEDV), the proportion of residential land zoned for lots over 25,000 sq.ft.(ZN), the crime rate(CRIM), the full-value property-tax rate(TAX), etc. The database consists of the data on 506 towns. We let $\mu = 0.2$, $\sigma = 0.2$, and $\gamma = 0.95$, and we define three linguistic terms on every attribute : **Z**ero, **M**edium, and **L**arge. The number of the discovered covering sets in the binary membership tuples is 639. The number of the attributes in the largest covering set is 8. We found 686 association rules and the followings are the parts of the discovered association rules.

- (CRIM, **L**) ⇒ (TAX, **L**)
- (MEDV, **Z**) ⇒ (CRIM, **Z**)
- (INDUS, **M**) (DIS, **Z**) (PTRATIO, **L**) ⇒ (TAX, **L**)
- (ZN, **Z**)(DIS, **Z**)(PTRATIO, **L**) (MEDV, **L**) ⇒ (INDUS, **L**)

The first and the second have simple forms and the others have more complicated forms. The first two rules are interesting. They say that if the crime rate is high, then the full-value property-tax rate is also high, and if the median value of owner-occupied homes is low then the crime rate is also low.

The third also tells about the tax. That is, the town where the non-trail business areas is middle-wide and the distance to the five employment centers are short and the pupil-teacher ratio is high, has the high tax ratio.

5 Concluding Remarks

In this paper, we have proposed a method finding association rules among real-valued attributes using fuzzy sets. By using fuzzy sets, we describe the association rule in a concise manner. The user can make effect on the discovering process by changing the number and the position of the fuzzy sets, and get the different kinds and different detail levels of the discovered rules.

It is clear that the proposed association rules are the extension of the existing association rules. The proposed method can discover and describe all that the existing method can do. If we define only one signleton representing '1' and apply the proposed method to binary data tuples, we can get the association rules which we can get from the existing method.

The number of discovered association rule may become so large that it would be difficult to use directly the rules themselves. Therefore, the research on redundancy removing or re-organizing the discovered rule are required.

References

[1] W.J. Frawley, G. Piatetsky-Shapiro, C.J. Matheus, *Knowledge Discovery in Databases: An Overview*, Knowledge Discovery in Databases(G. Piatetsky-Shapiro, W.J. Frawley, eds.), pp.1-27 AAAI Press/The MIT Press, US, 1993.

[2] D.H. Lee, *Data Mining Techniques : An Overview*, 23th KFIS Fall Conference, special lecture notes, pp.23-35, Seoul, Korea, 1996.

[3] M. Holsheimer, M. Kersten, H. Mannila, H. Toivonen, *A Perspective on Databases and Data Mining*, the 1st International Conference on Knowledge Discovery and Data Mining, pp.150-155, 1995.

[4] M. Flemettinen, H. Mannila, P. Ronkainen, H. Toivonen, A. I. Verkamo, *Finding Interesting Rules from Large Sets of Discovered Association Rules*, the 3rd International Conference on Information and Knowledge Management, pp.401-407, 1994.

[5] H. Mannila, H. Toivonen, A.I. Verkamo, *Efficient Algorithms for Discovering Association Rules*, AAAI workshop on Knowledge Discovery in Databases, pp.181-192, Washington, July, 1994.

[6] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A.I. Verkamo, *Fast Discovery of Association Rules*, Advances in Knowledge Discovery and Data Mining(U.M. Fayyad, et al. eds.), pp.307-328, AAAI Press/The MIT Press, US, 1996.

*) Merz, C.J., & Murphy, P.M. (1996). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn /MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.