Mining Generalized Association Rules with Fuzzy Taxonomic Structures

Qiang Wei*, Guoqing Chen[1]**
Division of Management Science & Engineering,
School of Economics & Management, Tsinghua University, Beijing 100084, China
*weiq@em.tsinghua.edu.cn
**chengq@em.tsinghua.edu.cn

Abstract

Data mining is a key step of knowledge discovery in databases. Usually, Srikant and Agrawal's algorithm is used for mining generalized association rules upon all levels of presumed exact taxonomic structures. However, in many real-world applications, the taxonomic structures may not be crisp but fuzzy. This paper focuses on the issue of mining generalized association rules with fuzzy taxonomic structures. Particular attention is paid to extending the notions of the degree of support, the degree of confidence, and the R-interest measure. The computation of these degrees takes into account the fact that there exists a partial belonging of any two itemsets in the taxonomy concerned. Finally, a simplified example is given to help illustrate the ideas.

1. Introduction

Data mining is a key step of knowledge discovery in large databases. One of the important issues in the fields is to efficiently discover the relationship among data items in forms of association rules that are of interest to decision-makers. In 1993, Agrawal [1] proposed an algorithm for mining association rules that represent the relationships between basic data items (e.g., items from original sales records). An example of such rules is Apple ⇒ Pork (e.g., customers who bought apples turned to buy pork). Further, various efforts have been made to improve or extend the algorithm, e.g., [2,4-9]. In Srikant and Agrawal [8], the algorithm is extended to allow the discovery of the so-called generalized association rules that represent the relationships between basic data items, as well as between data items at all levels of related taxonomic structures. An example of such rules is Fruit ⇒ Meat (e.g., customers who bought fruit products turned to buy meat products). Notably, the computation of the degree of support, denoted hereafter as Dsupport, and the degree of confidence, denoted hereafter as Dconfidence, plays an important role in the algorithm [8]. Specifically, Dsupport and Dconfidence of the generalized associated rule X⇒Y are defined as follows:

$$Dsupport(X \Rightarrow Y) = \|X \cup Y\| / |T|$$

$$Dconfidence(X \Rightarrow Y) = \|X \cup Y\| / \|X\|$$

Where X and Y are itemsets with X∩Y=∅, T is the set of transactions contained in the database concerned, ‖X‖ is the number of transactions in T that contain X, ‖X∪Y‖ is the number of transactions in T that contain X and Y, and |T| is the number of transactions contained in T. Mining

[1] Corresponding author

0-7803-5211-4/99/$10.00 © 1999 IEEE

generalized associated rules X⇒Y, if any, in a database is to find whether the transactions in the database satisfy the pre-specified thresholds, min-support and min-confidence, for Dsupport and Dconfidence respectively. Usually, the rules discovered in this way may need to be further filtered, for instance, to eliminate redundant and inconsistent rules using the R-interest measure, which will be discussed in Section 2.

However, in many real world applications, the related taxonomic structures may not be necessarily crisp, rather, certain fuzzy taxonomic structures reflecting partial belonging of one item to another may pertain. For example, Tomato may be regarded as being both Fruit and Vegetable, but to different degrees. An example of a fuzzy taxonomic structure is shown in Figure 1. Apparently, in such a fuzzy context, the computation of Dsupport and Dconfidence shown above can hardly be applied, but needs to be extended accordingly.

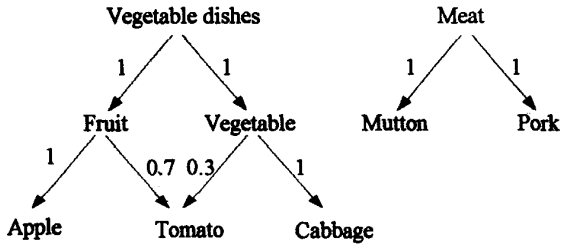

Figure 1 Example of fuzzy taxonomic structures

2. Fuzzy extensions

A crisp taxonomic structure assumes that the child item belongs to its ancestor with degree 1. But in a fuzzy taxonomy, this assumption is no longer true. Different degrees may pertain across all nodes (itemsets) of the structure.

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of literals, called items. Let FG be a directed acyclic graph (DAG) on the literals [10]. An edge in FG represents a fuzzy *is-a* relationship, which means along with each edge, there exists a partial degree μ with which the child-node on this edge belongs to its parent-node on this edge, where $0 \le \mu \le 1$. If there is an edge in FG from *p* to *c*, we call *p* a parent of *c* and *c* a child of *p* (*p* represents a generalization of *c*.). We model the fuzzy taxonomic structure as a DAG rather than a forest to allow for multiple taxonomies.

We call x^ an *ancestor* of x (and x a descendant of x^) if there is a directed path (a series of edges) from x^ to x in the transitive-closure of FG. Note that a node is not an ancestor of itself, since the graph is acyclic.

Let T be a set of all transactions, I be a set of all items, and t be a transaction in T such that $t \subseteq I$. Then, we say that a transaction t supports an item $x \in I$ with degree 1 if x is in t, or with degree μ, if x is an ancestor of some item y in t such that y belongs to x in a degree μ. We say that a transaction t supports $X \subseteq I$ with degree β:

$$\beta = \min_{x \in X}(\mu)$$

where μ is the degree of x in X that is supported by t, $0 \le \mu \le 1$.

Generally, given a transaction set T, there may exist a fuzzy taxonomic structure FG as shown in Figure 2.

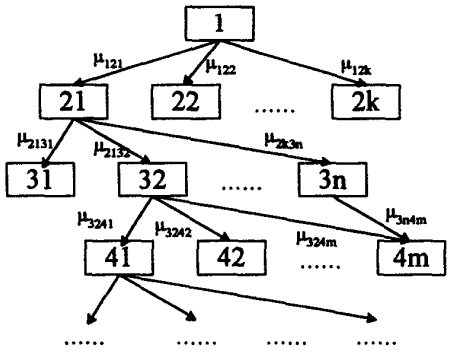

Figure 2 A fuzzy taxonomic structure

Every child-node x belongs to its parent-node y with degree μ_{yx}, $0 \le \mu_{yx} \le 1$. The leaf-nodes of the structure are attribute values of the transaction records. Every non-leaf-node is referred to as an attribute-node, which is regarded

as a set whose elements are the leaf-nodes with respective membership degrees. Sometimes for the purposes of convenience, each leaf-node is regarded as a set that contains merely the attribute value itself. In this case, the attribute value belongs to the leaf-node with a degree of 1. As the fuzzy taxonomic structure only represents the partial degrees of the edges, the degrees between leaf-nodes and attribute-nodes in FG need to be derived. This could be done based upon the notions of subclass, superclass and inheritance, which have been discussed in [3]. Specifically,

$$\mu_{xy} = \bigoplus_{\forall l: x \to y} \left(\bigotimes_{\forall e \text{on} l} \mu_{le} \right) \tag{1}$$

where l: $x \to y$ is one of the accesses (paths) of attributes x and y, e *on* l is one of the edges on access l, μ_{le} is the degree on the edge e on l. If there is no access between x and y, $\mu_{xy} = 0$. Notably, what specific forms of the operators to use for $\oplus$ and $\otimes$ depends on the context of the problems at hand. Merely for illustrative purposes, in this paper, ***max*** is used for $\oplus$ and ***min*** for $\otimes$.

Now consider the computation of the degree of support in such a fuzzy taxonomic structure case. If a is an attribute value in a certain transaction $t \in T$, T is the transaction set, and x is an attribute in certain itemset X, then the degree μ_{xa} with which a belongs to x can be obtained according to formula (1). Thus, μ_{xa} may be viewed as the degree that the transaction $\{a\}$ supports x. Further, the degree that t supports X can be obtained as follows:

$$\mu_{tX} = Support_{tX} = \min_{x \in X} (\max_{a \in t} (\mu_{xa})) \tag{2}$$

In this way, the degree that a transaction t in T supports a certain itemset X is computed. Moreover, in terms of how many transactions in T support X, the $\Sigma count$ operator [3] is used to sum up all the degrees that are associated with the transactions in T:

$$\sum_{\forall t \in T} count\,(Support_{tX}) = \sum_{\forall t \in T} count\,(\mu_{tX}) \tag{3}$$

Hence, for a generalized association rule X⇒Y, let X∪Y = Z ⊆ I, then Dsupport (X⇒Y) can be obtained as follows:

$$\sum_{\forall t \in T} count\,(\mu_{tZ}) / |T| \tag{4}$$

In an analogous manner, $Dconfidence(X \Rightarrow Y)$ can be computed as follows:

$$\sum_{\forall t \in T} count\,(\mu_{tZ}) / \sum_{\forall t \in T} count\,(\mu_{tX}) \tag{5}$$

Furthermore, the concept of R-interest [8] can be extended based on the notion of Dsupport. Like in the classical case, the extended R-interest measure is a way used to prune out those "redundant" rules. Briefly speaking, the rules of interest, according to R-interest, are those rules whose degrees of support are more than R times the expected degrees of support or whose degrees of confidence are more than R times the expected degrees of confidence.

Consider a rule X⇒Y, where $X=\{x_1, x_2, \ldots, x_m\}$ and $Y=\{y_1, y_2, \ldots, y_n\}$. X^ and Y^ are called the ancestors of X and Y respectively, if $X^=\{x^_1, x^_2, \ldots, x^_m\}$ where $x^_i$ is an ancestor of x_i, $1 \le j \le m$, and $Y=\{y^_1, y^_2, \ldots, y^_n\}$, where $y^_j$ is an ancestor of y_j, $1 \le j \le n$. Then the rules X^⇒Y, X^⇒Y^ and X⇒Y^ are called the ancestors of the rule X⇒Y. Given a set of rules, we call X^⇒Y^ a *close ancestor* of X⇒Y if there is no rule X'⇒Y' such that X'⇒Y' is an ancestor of X⇒Y and X^⇒Y^ is an ancestor of X'⇒Y'. Let $Dsupport_{E(X^\Rightarrow Y^)}(X \Rightarrow Y)$ denote the "expected" value of the degree of support of rule X⇒Y and $Dconfidence_{E(X^\Rightarrow Y^)}(X \Rightarrow Y)$ denote the "expected" value of the degree of confidence, then with fuzzy taxonomic structures, we have

$$\begin{aligned} & D\sup port_{E(X^\Rightarrow Y^)}(X \Rightarrow Y) \\ &= \sum count(\mu_{t\{x_1\}}) \Big/ \sum count(\mu_{t\{x^_1\}}) \times \ldots \times \\ & \sum count(\mu_{t\{x_m\}}) \Big/ \sum count(\mu_{t\{x^_m\}}) \times \\ & \sum count(\mu_{t\{y_1\}}) \Big/ \sum count(\mu_{t\{y^_1\}}) \times \ldots \times \\ & \sum count(\mu_{t\{y_n\}}) \Big/ \sum count(\mu_{t\{y^_n\}}) \times \\ & D\sup port(X^\Rightarrow Y^) \end{aligned} \tag{6}$$

and

$$Dconfidence_{E(X^{\wedge}\Rightarrow Y^{\wedge})}(X \Rightarrow Y)$$
$$= \sum count(\mu_{t\{y_1\}}) / \sum count(\mu_{t\{y^{\wedge}_1\}}) \times \ldots \times$$
$$\sum count(\mu_{t\{y_n\}}) / \sum count(\mu_{t\{y^{\wedge}_n\}}) \times \quad (7)$$
$$Dconfidence(X^{\wedge} \Rightarrow Y^{\wedge})$$

Notably, in the case of crisp taxonomic structures, $\Sigma count(\mu_{t\{xi\}})$ and $\Sigma count(\mu_{t\{yi\}})$ degenerate to $\|\{x_i\}\|$ and $\|\{y_i\}\|$ respectively.

3. An example

Suppose that a supermarket maintains a database for the goods that customers have purchased as shown in Table 1.

Transaction #	Something Bought
#100	Apple
#200	Tomato, Mutton
#300	Cabbage, Mutton
#400	Tomato, Pork
#500	Pork
#600	Cabbage, Pork

Table 1 Transactions in a supermarket database

The *min-support* is 30% (2 transactions), the *min-confidence* is 60%, the R-interest is 1.2. It is assumed that the underlying taxonomic structures are fuzzy and as shown in Figure 1. Then, according to formula (1) we have Table 2 for those leaf-nodes and their ancestor's degrees. For instance, in Table 2, μ(Tomato∈Vegetable dishes) = max(min(1,0.7), min(1, 0.3)) = 0.7.

Leaf-nodes	The degrees of the ancestors and its own
Apple	1/Apple, 1/Fruit, 1/Vegetable dishes
Tomato	1/Tomato, 0.3/Vegetable, 0.7/Fruit, 0.7/Vegetable dishes
Cabbage	1/Cabbage, 1/Vegetable, 1/Vegetable dishes
Pork	1/Pork, 1/Meat
Mutton	1/Mutton, 1/Meat

Table 2 Leaf-nodes and their ancestor's degrees

Furthermore, according to the formula (3) for the $\Sigma count$ values, all the frequent itemsets are listed in Table 3 along with their corresponding $\Sigma count$ values. Here, by a frequent itemset we mean the itemset whose degree of support is more than the *min-support*.

Frequent Itemsets	$\Sigma count$ values
{Cabbage}	2
{Tomato}	2
{Pork}	3
{Mutton}	2
{Fruit}	2.4
{Vegetable}	2.6
{Vegetable dishes}	4.4
{Meat}	5
{Cabbage, Meat}	2
{Tomato, Meat}	2
{Vegetable, Meat}	2.6
{Vegetable dishes, Meat}	3.4

Table 3 $\Sigma count$ values for frequent itemsets

In Table 3, the $\Sigma count$ value for the itemset {Vegetable, Meat} is calculated as:

min(0.3, 1) + min(1, 1) + min(0.3, 1) +min(1, 1) = 2.6

Based on these $\Sigma count$ values for all the frequent itemsets, the degrees of support for all possible rules can be computed. Table 4 lists those rules discovered, which satisfy the given thresholds 30%, 60%, 1.2 for the degree of support, the degree of confidence, and the R-interest, respectively. For instance, *Dsupport*(Vegetable dishes ⇒ Meat) = 3.4/6 = 57%, and *Dconfidence*(Vegetable dishes ⇒ Meat) = 3.4/4.4 = 77%. It is worth mentioning that the rule Cabbage⇒Meat is filtered out, though with *Dsupport*(Cabbage ⇒ Meat) = 2/6 = 33% > 30% and *Dconfidence*(Cabbage ⇒ Meat) = 2/2 = 100% > 60%. This is done according to the R-interest measure:

$$Dsupport_{E(Vegetable \Rightarrow Meat)}(Cabbage \Rightarrow Meat)$$
$$= \Sigma count(\mu_{t(Cabbage)}) / \Sigma count(\mu_{t(Vegetable)}) \times$$
$$\Sigma count(\mu_{t(Meat)}) / \Sigma count(\mu_{t(Meat)}) \times$$
$$Dsupport(Vegetable \Rightarrow Meat)$$
$$= 2/2.6 \times 5/5 \times 2.6/6$$
$$= 33\%$$

$Dconfidence_{E(Vegetable \Rightarrow Meat)}(Cabbage \Rightarrow Meat)$
$= \sum count(\mu_{t(Meat)}) / \sum count(\mu_{t(Meat)}) \times$
$Dconfidence(Vegetable \Rightarrow Meat)$
$= 5/5 \times 100\%$
$= 100\%$

and Dsupport(Cabbage⇒Meat) / $Dsupport_{E(Vegetable \Rightarrow Meat)}$ (Cabbage⇒Meat) = 33% / 33% = 1.0 < 1.2, and Dconfidence(Cabbage⇒Meat) / $Dconfidence_{E(Vegetable \Rightarrow Meat)}$ (Cabbage⇒Meat) = 100% / 100% = 1.0 < 1.2, which means that this rule is regarded redundant with respect to the existing rule "Vegetable⇒Meat" in Table 4.

Interesting Rules	Dsupport	Dconfidence
Vegetable⇒Meat	43%	100%
Vegetable dishes⇒Meat	57%	77%
Meat⇒Vegetable dishes	57%	68%

Table 4 The discovered rules of interest

4. Conclusions

Aimed at dealing with the taxonomic inexactness when mining association rules, this paper has introduced the fuzziness in the underlying taxonomic structures and extended the classical algorithm in a way that a transaction may partially support a particular item. This has then led to re-examining the computation for the degree of support and the degree of confidence, as well as for the R-interest measure. An example has been given to illustrate the proposed fuzzy extensions. Further studies are being conducted to consider more general forms of association rules, e.g., fuzzy association rules, and to develop corresponding computational algorithms.

References

1 Rakesh Agrawal, Tomasz Imielinski, Arun Swami, *Mining Association Rules between Sets of Items in Large Databases*, Proceedings of the 1993 ACM SIGMOD Conference Washington DC, USA, May 1993.

2 Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, A. Inkeri Verkamo, *Fast Discovery of Association Rules* in Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press, 1996.

3 Guoqing Chen, Fuzzy Logic in Data Modeling: semantics, constraints and database design, Kluwer Academic Publishers, Boston, 1998.

4 J, Han, Y. Fu, *Discovery of Multiple-level Association Rules from Large Databases*, Proceedings of the 21st International Confernce on Very Large Databases, Zurich, Switzerland, September 1995.

5 Takeshi Fukuda, Yasuhiko Morimoto, Shinichi Morishita, Takeshi Tokuyama, *Data Mining Using Two Dimensional Optimized Association Rules: Scheme, Algorithm and Visualization*, SIGMOD'96 6/96 Montreal Canada, 1996.

6 Maurice Houtsma, Arun Swami, *Set-oriented Data Mining in Relational Databases*, Data & Knowledge Engineering, 17 (1995) 245-262.

7 Savasere, E. Omiecinski, S. Navathe. *An Efficient Algorithm for Mining Association Rules in Large Databases*, Proceedings of the VLDB Conference, Zurich, Switzerland, September 1995.

8 Ramakrishnan Srikant, Rakesh Agrawal, *Mining Generalized Association Rules*, Proceedings of the 21st VLDB Conference Zurich, Swizerland, 1995.

9 Ramakrishnan Srikant, Rakesh Agrawal, *Mining Quantitative Association Rules in Large Relational Tables*, SIGMOD'96 6/96 Montreal, Canada, 1996.

10 Weimin Yan, Weimin Wu, Data Structure, Tsinghua University Press, 1992.