

Multidimensional Time Series Fuzzy Association Rules Mining

Xuedong Gao

Hongwei Guo

School of Management,

University of Science and Technology Beijing, P.R.China, 100083

gaoxuedong@manage.ustb.edu.cn

ABSTRACT

In this paper, we present a new solution, in which the fuzziness of both subsequences and subsequences interval has been taken into consideration for solving the problem of multidimensional time series fuzzy association rules mining. Aimed at dealing with the new conception, this paper has put forward some key algorithms of the solution. Finally, an application example of multidimensional time series fuzzy association rules mining is illustrated. The result shows that rules with fuzzy interval can only be mined out by the above-mentioned new method.

Key words: Data mining, Time series, Fuzzy association rules

INTRODUCTION

Association rules reflect the relationships among data items that are of interest to decision-makers (Chen & Wei, 2002). In 1993, Agrawal, Imicliniski and Swami (1993) proposed an approach to mining association rules that represent the relationships among basic data items. The time series data is a set of data sequences, each of which is an ordered list of elements (Kim, Park & Lee, 2004). So the core of the time series association rules mining is to find the relationships among the subsequences.

Generally, the research on time series association rules mining considers the subsequence pattern as crisp property. Das et al. researched the time series crisp association relation on local patterns (Das, Lin & Mannila, 1998). Mark Last et al. researched the time series crisp association relationships between subsequence pattern and future trend (Mark, Klein & Kandel, 2001; Jiang & Cai, 2003; Dong, Shi & Li, 2003; Wang, Zhang & Gao, 2005). Since the taxonomies of subsequences may not be crisp but fuzzy in fact (e.g., “the growth rate of one stock is 3% in five days” could be expressed in linguistic terms as: “almost no increase” and “increase slowly”, each has different degree of membership.), recent research imports the notion of fuzzy set (Chiang, Chow & Wang, 2000; Au & Chan, 2005). For instance, WANG Binxue put forward a new algorithm of time series fuzzy association rules mining which had considered the fuzziness of subsequences (Wang, 2004; Zhang, Zhang & Chen, 2002). But there are still two limitations on their research. Firstly, they ignored the fuzziness of subsequences interval. Secondly, time span is not fit for rules description especially on multidimensional time series fuzzy association rules mining.

The work stems from the two limitations above. Based on the fuzziness of both subsequences and subsequences interval, a newly improved algorithm of multidimensional time series association rules mining has been put forward with the application of fuzzy set mathematical theory.

METHODOLOGY

Time Series Fuzziness and Discretization

First, the time series need the treatment of fuzziness and discretization in the process of time series association rules mining. It is not reasonable to cluster the subsequences crisply. We shall cluster every subsequence in terms of the fuzzy set theory.

Formally, let $s = (x_1, x_2, \dots, x_N)$ be a time series, putting a time window W in width on the series forms the

subsequence $s_i = (x_i, x_{i+1}, \dots, x_{i+w-1})$. With the time window one-step sliding from the start to the end, a series of subsequences $x_1, x_2, \dots, x_{N-w+1}$ with w in width forms. The collection of the series of subsequences can be expressed as follows:

$$W(s, w) = \{s_i \mid i = 1, 2, \dots, N - w + 1\} \tag{1}$$

(1) Regard $w(s, w)$ as one of $N - w + 1$ subsequences in w dimensional Euclidean space, and cluster these subsequences into k clusters randomly. Calculate every cluster's center, the coordinates l of the cluster j can be expressed as:

$$x_{j,l} = \frac{1}{h} \sum_{i=1}^h x_{j,l,i}, l = 1, \dots, w, j = 1, \dots, k \tag{2}$$

(2) Take the center of every cluster as its representative subsequence, the element s_i in the collection $w(s, w)$ belongs to the cluster j with the $\mu_j(s_i)$ degree of membership, then calculate every elementary degree of membership. The Membership function $\mu_j(s_i)$ can be expressed as follows:

$$\mu_j(s_i) = \frac{\left(\frac{1}{\|s_i - x_j\|^2}\right)^{\frac{1}{b-1}}}{\sum_{c=1}^k \left(\frac{1}{\|s_i - x_c\|^2}\right)^{\frac{1}{b-1}}}, j = 1, \dots, k \tag{3}$$

Where $b > 1$ and b is a constant which can control the fuzzy degree of cluster result, $\|s_i - x_j\|^2$ represents the square of the distance between every subsequence and the cluster j .

(3) Use current membership function to calculate the center of every cluster again. It can be expressed as follows:

$$x_{j,l} = \frac{\sum_{i=1}^{(N-w+1)} [\mu_j(s_i)]^b x_{j,l,i}}{\sum_{i=1}^{(N-w+1)} [\mu_j(s_i)]^b} \tag{4}$$

$j = 1, 2, \dots, k; l = 1, 2, \dots, w$

Repeat the calculation of the step (2) and (3) above until every sample's degree of membership becomes stable. The collection of representative subsequences can be expressed in the form of $D = \{x_1, x_2, \dots, x_k\}$, where x_j represents the cluster j representative subsequence.

(4) Fuzziness of subsequences interval

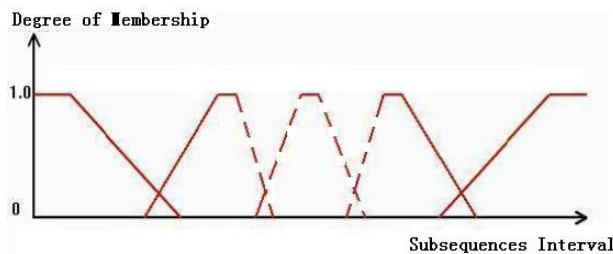


Figure.1: Membership function of subsequences interval

Select Fig.1 as the membership function of subsequences interval or the other compatible membership function. The result of membership function can be expressed in the form of $u_T(N)$, where N is subsequences interval. The collection of fuzzy subsequences interval can be expressed as $Q = \{t_1, t_2, \dots, t_f\}$, where t_i represents the state i of fuzzy subsequences interval, and f represents the total number of fuzzy subsequences interval states.

2.2 The description and selection of fuzzy association rules

This paper focuses on fuzzy association rules that are in the form of “if A happens, then B will happen before long”, where linguistic term “before long” can be represented by T , so the rule can be expressed in the form of $A \Rightarrow (T, B)$, where $A, B \in \{x_1, x_2, \dots, x_k\}, T \in \{t_1, t_2, \dots, t_f\}$. The frequency of that A happens can be defined as follows:

$$F(A) = \sum_{i=1}^{(N-w+1)} u_A(s_i) \tag{5}$$

Wherein, subsequence s_i belongs to the representative subsequence A with the $u_A(s_i)$ degree of membership.

To the fuzzy rule $A \Rightarrow (T, B)$, its degree of confidence (Dconf) can be expressed as follows:

$$c(A \Rightarrow (T, B)) = \frac{F(A, B, T)}{F(A)} \tag{6}$$

Wherein, $F(A, B, T)$ represents the frequency of the rule, and it can be calculated as follows:

$$F(A, B, T) = \sum_{i=1}^{(N-w)} \{u_A(s_i) \wedge \sum_{j=i+1}^{N-w+1} [u_B(s_j) \wedge u_T(j-i)]\} \tag{7}$$

The degree of support reflects the rule’s statistical significance, while Dconf is a measure of rule’s strength. The problem of mining association rules is to find all strong association rules whose Dsupp and Dconf are no less than the pre-specified minimal support and minimal confidence, respectively. In the paper, the rule’s degree of lift (Dlift) is also been considered. The Dlift is calculated as follows:

$$lift(A \Rightarrow (T, B)) = \frac{c(A \Rightarrow (T, B))}{p(T, B)} \tag{8}$$

Wherein, $p(T, B)$ is the frequency of that subsequence B is T interval to subsequence A, and it can be expressed as follows:

$$p(T, B) = \sum_{i=1}^{(N-w)} u_A(s_i) \wedge \frac{\sum_{j=i+1}^{N-w+1} [u_B(s_j) \wedge u_T(j-i)]}{N-w+1} \tag{9}$$

Multidimensional time series fuzzy association rules mining

To a m dimensional time series, the collection of subsequences can be obtained through the treatment of sliding time window. The collection of subsequences can be expressed as follows:

$$W(s, w) = \{s_i^h | i = 1, 2, \dots, (N-w+1); h = 1, 2, \dots, m\}$$

After all the m sub-collections $\{s_i^h | i = 1, 2, \dots, (N-w+1)\}, h = 1, 2, \dots, m$ of $w(s, w)$ are treated by fuzziness and discretization, every sub-collection has k representative subsequences. The collection of representative subsequences of every sub-collection can be expressed as:

$$D^h = \{x_1^h, x_2^h, \dots, x_k^h\}, h = 1, 2, \dots, m$$

Wherein, x_j^h represents the representative subsequence j of the sub-collection h . The collection of fuzzy subsequences interval can be expressed in the form of $Q^h = \{t_1, t_2, \dots, t_f\}$, $h = 1, 2, \dots, m$.

Define the form of a multidimensional fuzzy association rule as: “if A^1, A^2 happens with Q^1 interval, ..., A^p happens with Q^{p-1} interval, ..., A^h happens with Q^{h-1} interval, then B happens with Q^h interval”, the rule can be expressed in the following form: $A^1 \wedge Q^1 \wedge A^2 \wedge Q^2 \wedge \dots \wedge A^p \wedge Q^p \wedge \dots \wedge A^h$

$\Rightarrow (Q^h, B)$. Where $A^p \in D^p$, and $D^p \in \{D^1, D^2, \dots, D^m\}$, $p = 1, 2, \dots, h$, $B \in D^r$, $D^r \in \{D^1, D^2, \dots, D^m\}$, A^p is ordered by the sequence of subsequences, and $i, j = 1, 2, \dots, h, i \neq j$, $D^i \cap D^j = \emptyset$. The frequency of the rule’s antecedent can be defined as follows:

$$F(A^1 \wedge Q^1 \wedge A^2 \wedge Q^2 \wedge \dots \wedge A^p \wedge Q^p \wedge \dots \wedge A^h) = \sum_{i=1}^{N-w} u_{A^1}(s_i^1) \wedge NE(1,2) \tag{10}$$

$$NE(a_z, b) = \sum_{a_{z+1}=a_z+1}^{N-w+1} \{u_{A^b}(s_{a_{z+1}}^b) \wedge u_{Q^{b-1}}(a_{z+1} - a_z) \wedge NE(a_{i+1}, b+1)\} \tag{11}$$

Where $a_1 = 1$, $z < h, b < h$, when $z = h, b = h$

$$NE(a_h, h) = \sum_{a_h=a_{h-1}+1}^{N-w+1} [u_{A^h}(s_{a_h}^h) \wedge u_{Q^{h-1}}(a_h - a_{h-1})] \tag{12}$$

The Dconf of the fuzzy rule $A^1 \wedge Q^1 \wedge \dots \wedge A^p \wedge Q^p \wedge \dots \wedge A^h \Rightarrow (Q^h, B)$ is calculated as follows:

$$c(A^1 \wedge Q^1 \wedge \dots \wedge Q^p \wedge \dots \wedge A^h \Rightarrow (Q^h, B)) = \frac{F(A^1 \wedge Q^1 \wedge \dots \wedge Q^p \wedge \dots \wedge A^h; Q^h, B)}{F(A^1 \wedge Q^1 \wedge \dots \wedge Q^p \wedge \dots \wedge A^h)} \tag{13}$$

Where $F(A^1 \wedge Q^1 \wedge \dots \wedge Q^p \wedge \dots \wedge A^h; Q^h, B)$ represents the rule’s frequency, it can be calculated as follows:

$$F(A^1 \wedge Q^1 \wedge A^2 \wedge Q^2 \wedge \dots \wedge A^p \wedge Q^p \wedge \dots \wedge A^h; Q^h, B) \tag{14}$$

$$= \sum_{i=1}^{N-w} u_{A^1}(s_i^1) \wedge TE(1,2) \tag{15}$$

$$TE(a_z, b) = \sum_{a_{z+1}=a_z+1}^{N-w+1} [u_{A^b}(s_{a_{z+1}}^b) \wedge u_{Q^{b-1}}(a_{z+1} - a_z) \wedge TE(a_{i+1}, b+1)]$$

Where $a_1 = 1$, $z < h, b < h$, when $z = h, b = h$:

$$TE(a_h, h) = \sum_{a_h=a_{h-1}+1}^{N-w+1} \{u_{A^h}(s_{a_h}^h) \wedge u_{Q^{h-1}}(a_h - a_{h-1})\} \sum_{a_{h+1}}^{N-w+1} [u_b(s^r) \wedge u_{Q^h}(a_{h+1} - a_h)] \quad (16)$$

An application example

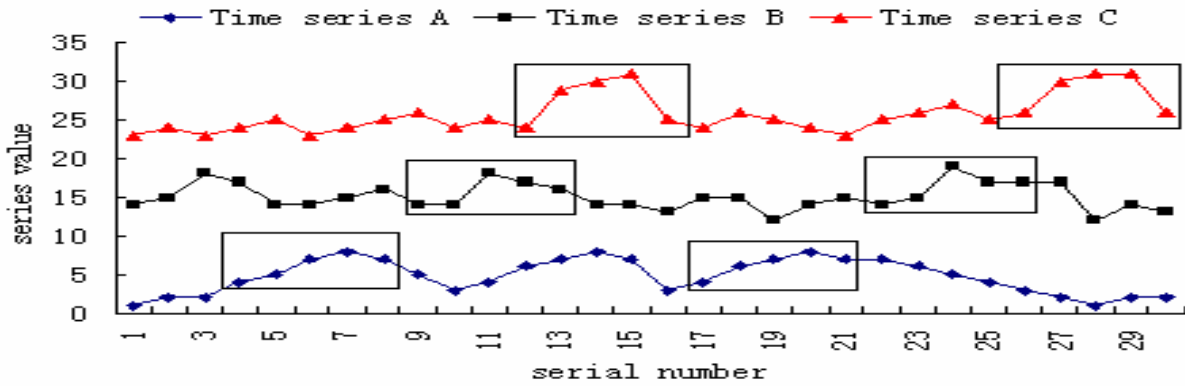


Fig.2 Three sets of time series

Time series	Cluster type	The degree of membership (Dmemb)
A	3	$u_{A3}(s_4) = 0.869, u_{A3}(s_{10}) = 0.861,$ $u_{A3}(s_{17}) = 0.841$
B	1	$u_{B1}(s_1) = 0.629, u_{B1}(s_9) = 0.741, u_{B1}(s_{22}) = 0.684$
C	4	$u_{C4}(s_{12}) = 0.913, u_{C4}(s_{26}) = 0.916$
Interval	2	$u_{T2}(3) = 0.8, u_{T2}(4) = 0.6$
	3	$u_{T3}(4) = 0.6, u_{T3}(5) = 0.85, u_{T3}(6) = 0.92,$ $u_{T3}(7) = 0.85, u_{T3}(8) = 0.6$

Table 1: Part of Dmemb

Rules	Dconf	Dsupp	Dlift
$A3 \Rightarrow (T3, A3)$	0.6927	0.0685	10.2
$B1 \Rightarrow (T3, B1)$	0.438	0.0527	18.98
$A3 \Rightarrow (T3, B1)$	0.537	0.0989	10.11
$B1 \Rightarrow (T2, C4)$	0.682	0.079	12.65
$(A3, T3, B1) \Rightarrow (T2, C4)$	0.941	0.0548	17.47

Table 2: The obtained rules

Suppose that there are three sets of time series that are illustrated in Fig.1. First, the Dmemb can be obtained through the process of time series fuzziness and discretization. Part of the result is shown in Tab.1, while some other values

of Dmemb that are related to the mining but less than 0.3 are supposed to be zero. Furthermore, the values of Dconf, Dsupp and Dlift are calculated with the corresponding algorithm given in the paper, and the result is shown in Tab.2. Finally, the rule $(A3, T3, B1) \Rightarrow (T2, C4)$ is chosen according to the pre-specified minimal Dconf, Dsupp and Dlift, respectively.

CONCLUSION AND FUTURE STUDIES

In this paper, we present a new solution, in which the fuzziness of both subsequences and subsequences interval has been taken into consideration for solving the problem of multidimensional time series fuzzy association rules mining. Aimed at dealing with the new conception, this paper has put forward some key algorithms of the solution. Finally, an application example of multidimensional time series fuzzy association rules mining is illustrated. The result shows that rules with fuzzy interval can only be mined out by the above-mentioned new method.

While applying the new method, there are still some aspects that shall be noted as follows:

(1) Both the value of time window w in width and the number of cluster k have an effect on the obtained result. Generally, when the research relates to time series short pattern, the value of w shall be small; otherwise it shall be big. The number of cluster k can be confirmed through analysing time series cluster result and merging some clusters to reduce unnecessary clusters.

(2) Appropriate minimal Dsupp and minimal Conf shall be established according to the number of antecedent variables. When the number of antecedent variables is bigger, the minimal Dsupp shall be smaller and the minimal Conf shall be bigger.

(3) Membership function of subsequences interval shall be established according to expert's knowledge. The appropriate membership function of subsequences interval not only can mining out effective association rules, but also can reduce the complexity of data mining.

REFERENCES

- Guoqing Chen, Qiang Wei, (2002). Fuzzy association rules and the extended mining algorithms, *Information Sciences*, vol. 147, pp.201–228,
- R. Agrawal, T. Imicliniski, & A. Swami, (1993). Mining association rules between sets of items in large databases, in *Proceedings of the 1993 ACM SIGMOD conference*, Washington, DC, USA, May.
- Sang-Wook Kim, Dae-Hyun Park, & Heon-Gil Lee, (2004). Efficient processing of subsequence matching with the Euclidean metric in time-series databases, *Information Processing Letters*, vol.90, pp.253–260.
- Das G., Lin K, & Mannila H (1998). Rule discovery from time series, in *Proceedings of the 3rd International Conference of Knowledge Discovery and Data Mining*, pp.16–22.
- Last Mark, Yaron Klein, Abraham Kandel, (2001). Knowledge Discovery in Time series Databases, in *IEEE transactions on systems, Man and cybernetics-part B: cybernetics*.
- Jiang Xiaoliang, Cai Zhihua, (2003). The research of data mining in time-series databases, *Microcomputer development (in Chinese)*, vol.13, pp.90–92, May.
- Dong Zekun, Shi Zhongzhi, Li Hui, (2003). An efficient algorithm for mining inter-time series association rules, *Computer engineer and application (in Chinese)*, vol.25, pp.196–198.
- Wang Yong, Zhang Xinzheng, Gao Xiangjun, (2005). Time series rules discovery and its algorithm, *Computer application research (in Chinese)*, vol.6, pp.23–24.
- Wai-Ho Au, Keith C.C. Chan, (2005). Mining changes in association rules: a fuzzy approach, *Fuzzy sets*

and systems, vol.149, pp.87–104.

Ding-An Chiang, Louis R. Chow, Yi-Fan Wang, Mining time series data by a fuzzy linguistic summary system, *Fuzzy sets and systems*, vol.112, pp.419–413, 2000.

Wang, Binxue, (2004). Time series fuzzy association rules mining, *Computer engineer and application (in Chinese)*, vol.12, pp.177–179.

Zhang, Xiaogang, Zhang, Jing, & Chen, Hua, (2002). Applying fuzzy time-series data mining for the fuzzy modeling of complex system, *Control theory and Applications (in Chinese)*, vol.6, pp.872–876.

