Temporal Fuzzy Association Rules Mining Based on Fuzzy Information Granulation

Zebang Li, Fan Bu, Fusheng Yu

Beijing Normal University
School of Mathematical Science

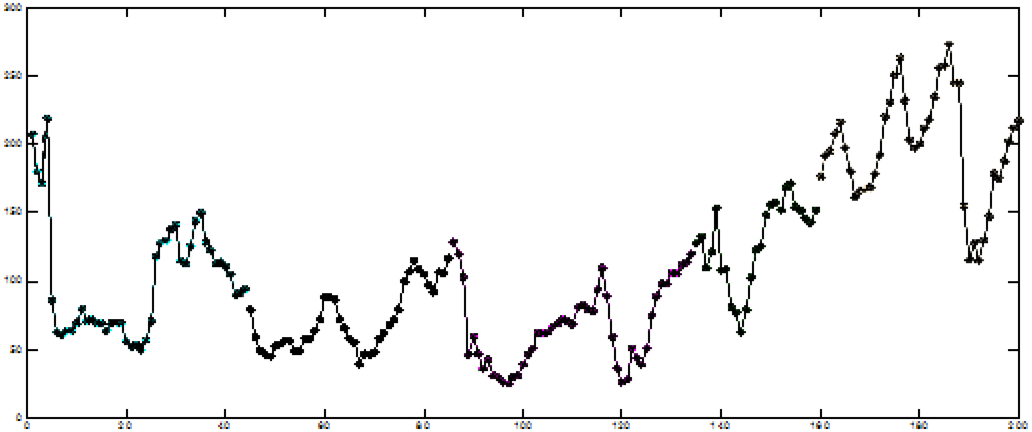
zebang@mail.bnu.edu.cn

July 28, 2017

Contents

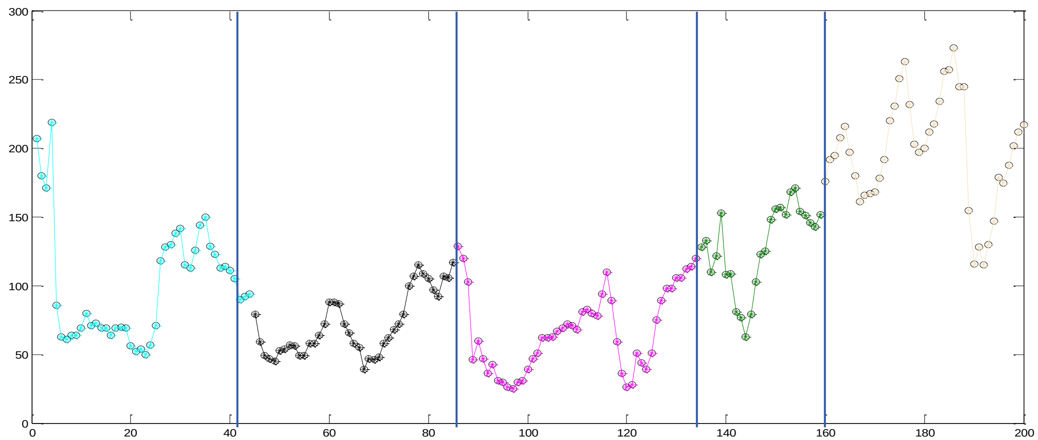
Introduction

Work Flow of Our Work



Introduction

Work Flow of Our Work



$$
\begin{array}{cc}
 & \begin{array}{ccccc} y_1 & y_2 & y_3 & \cdots & y_m \end{array} \\
\begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_N \end{array} & \begin{pmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1m} \\ u_{21} & u_{22} & u_{23} & \cdots & u_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ u_{n1} & u_{n2} & u_{n3} & \cdots & u_{nm} \end{pmatrix}
\end{array}
$$

Fuzzy Information Granulation

Fuzzy C-means for granular series

Temporal Fuzzy Association Rule Mining

Prerequisites

Association Rule

Association Rule Mining forms an important research area in the field of data mining.

Association Rule $A \rightarrow B$: If A occurred then B will occur.

Temporal Association Rule $A \xrightarrow{T} B$: If A occurred then B will occur after T.

Fuzzy Association Rule: A and B are fuzzy items.

Prerequisites

Fuzzy Information Granule

Gaussian Fuzzy Information Granule

$$f(x;\mu,\sigma) = \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where μ and σ represent the center(core) and spread of this fuzzy number.

Linear Gaussian Fuzzy Information Granule

$$f(x;kt+b,\sigma) = \exp\left(-\frac{\big(x-(kt+b)\big)^2}{2\sigma^2}\right), \qquad t \in [0,T],$$

where $\mu(t) = kt + b$ is a time-dependent core line, $k, b \in \boldsymbol{R}$ represent the slope and intercept of the core line respectively.

Prerequisites

Fuzzy C-Means for Granular Time Series

A finite collection of N Granulars is described as $T = \{t_1, t_2, t_3, \cdots, t_N\}$

and collection of m cluster centers is denoted $Y = \{y_1, y_2, \cdots, y_m\}$.

The fuzzy partition matrix is U,

$$\begin{array}{cc} & \begin{array}{ccccc} y_1 & y_2 & y_3 & \cdots & y_m \end{array} \\ \begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_N \end{array} & \begin{pmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1m} \\ u_{21} & u_{22} & u_{23} & \cdots & u_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ u_{n1} & u_{n2} & u_{n3} & \cdots & u_{nm} \end{pmatrix} = U. \end{array}$$

where $u_{ij} = t_i(y_j)$ is the membership degree of granule t_i to cluster y_j, $i \in \{1,2, \cdots, N\}$, $j \in \{1,2, \cdots, m\}$.

Prerequisites

Support Rate of Fuzzy Association Rule

Table: Salary Database and Fuzzy Clustering

Salary	Fuzzy Cluster		
	High	Middle	Low
S1=5000	0.21	0.29	0.50
S2=15000	0.41	0.41	0.18
S3=10000	0.26	0.48	0.26
S4=20000	0.52	0.41	0.07
S5=2000	0.08	0.17	0.75

$T = \{S1, S2, S3, S4, S5\}$ is a transaction set of salary.
$Y = \{"High", "Middle", "Low"\}$ is the fuzzy cluster.

For any sub set of Y, $Y' = \{y_1, y_2, \cdots, y_p\}$, $y_i \in Y$, the fuzzy support rate of Y' is defined as

$$sup(Y') = \frac{\sum_{j=1}^{n} \prod_{m=1}^{p} t_j(y_m)}{n},$$

where n and p are the number of elements in transaction set T and item set Y'.

Prerequisites

Support Rate of Fuzzy Association Rule

Table: Salary Database and Fuzzy Clustering

Salary	Fuzzy Cluster		
	High	Middle	Low
S1=5000	0.21	0.29	0.50
S2=15000	0.41	0.41	0.18
S3=10000	0.26	0.48	0.26
S4=20000	0.52	0.41	0.07
S5=2000	0.08	0.17	0.75

Then fuzzy support rate of association rule $Y_1 \to Y_2$ is,

$$sup(Y_1 \to Y_2) = \frac{\sum_{j=1}^{n} \prod_{m=1}^{p+q} t_j(y_m)}{n}$$

In this case, $n = 5$ and $p = 3$.

Prerequisites

Support Rate of Fuzzy Association Rule

Table: Temporal Salary Database and Fuzzy Clustering

Salary	Fuzzy Cluster		
	High	Middle	Low
T1=5000	0.21	0.29	0.50
T2=15000	0.41	0.41	0.18
T3=10000	0.26	0.48	0.26
T4=20000	0.52	0.41	0.07
T5=2000	0.08	0.17	0.75

In most studies, fuzzy association rules are not focused on time sequences.

Traditional association rules are no longer applicable to temporal data.

Now, for time series, each row of Table is not independent any more. The order of rows in Table represents the order of time.

此页照念即可

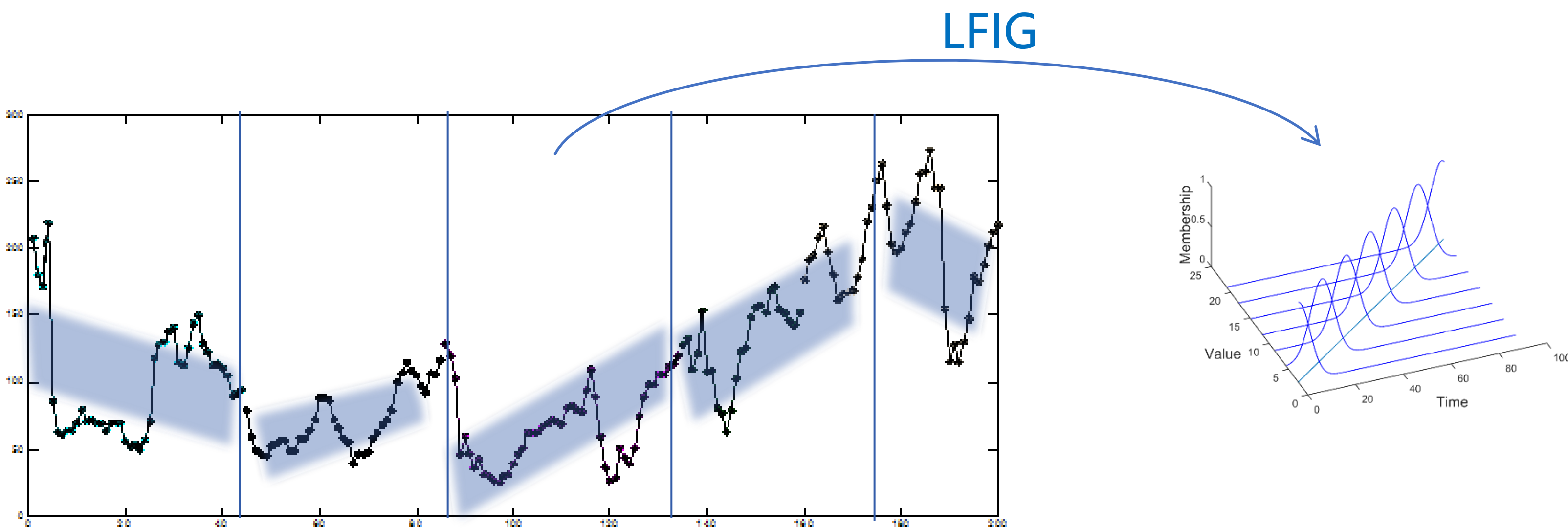
Temporal Fuzzy Association Rule

In this section

We extended the definition of fuzzy association rule from none-sequential to sequential.

We showed that our definition has a very low computation complexity.

Temporal Fuzzy Association Rule



For original time series $T = \{x_1, x_2, x_3, \cdots, x_N\}$, the granular time series $T' = \{t_1, t_2, t_3, \cdots, t_{N-l}\}$ is obtained by equal size granulation of LFIG.

Temporal Fuzzy Association Rule

$$
\begin{array}{cc}
 & \begin{array}{ccccc} y_1 & y_2 & y_3 & \cdots & y_m \end{array} \\
\begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_N \end{array} &
\begin{pmatrix}
u_{11} & u_{12} & u_{13} & \cdots & u_{1m} \\
u_{21} & u_{22} & u_{23} & \cdots & u_{2m} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
u_{n1} & u_{n2} & u_{n3} & \cdots & u_{nm}
\end{pmatrix} = U.
\end{array}
$$

Granular time series $T' = \{t_1, t_2, t_3, \cdots, t_{N-l}\}$ is clustered by fuzzy c-means. Based on FCM clustering, partition matrix U is obtained.

Temporal Fuzzy Association Rule

Definition 3: For $Y' = \{y_{i_1}, y_{i_2}, \cdots, y_{i_p}\}$, fuzzy support rate of Y' is defined as:

$$sup(Y) = \frac{\sum_{k=0}^{n-p} \prod_{j=1}^{j=p} u_{j+k,i_j}}{n-p},$$

where n is total number of granules, $u_{ij} = t_i(y_j)$ is membership degree of granule t_i to cluster y_j.

Association rule learning typically does not consider the order of clusters either within a transaction or across transactions. For time series, each row of the partition matrix U is not independent any more. The order of rows in matrix U represents the order of time. So in our definition, multiplications of memberships are carried out of each row by time order.

Temporal Fuzzy Association Rule

Example - Traditional

Price	Fuzzy Attribute		
	High	Middle	Low
S1=5000	0.21	0.29	0.50
S2=15000	0.41	0.41	0.18
S3=10000	0.26	0.48	0.26
S4=20000	0.52	0.41	0.07
S5=2000	0.08	0.17	0.75

$T = \{S_1, S_2, S_3, S_4, S_5\}$ is none-sequential dataset.

$Y = \{"High", "Middle", "Low"\}$ is fuzzy cluster.

Fuzzy support rate of $Y' = \{y_1, y_2, \cdots, y_p\}$, $y_i \in Y$, defined as:

$$sup(Y') = \frac{\sum_{j=1}^{n} \prod_{m=1}^{p} S_j(y_m)}{n}$$

Temporal Fuzzy Association Rule

Example - Traditional

Price	Fuzzy Attribute		
	High	Middle	Low
S1=5000	0.21	0.29	0.50
S2=15000	0.41	0.41	0.18
S3=10000	0.26	0.48	0.26
S4=20000	0.52	0.41	0.07
S5=2000	0.08	0.17	0.75

Support rate of $Y' = \{High, Middle\}$ is

$$sup(Y') = \frac{0.21*0.29+0.41*0.41+\ldots+0.08*0.17}{5}$$

Temporal Fuzzy Association Rule

Example - Temporal

Price	Fuzzy Attribute		
	High	Middle	Low
T1=5000	0.21	0.29	0.50
T2=15000	0.41	0.41	0.18
T3=10000	0.26	0.48	0.26
T4=20000	0.52	0.41	0.07
T5=2000	0.08	0.17	0.75

Fuzzy pattern $Y' = \{High, Middle\}$ means:

The first 'day' is High and the second 'day' is Middle.

Temporal Fuzzy Association Rule

Example - Temporal

Price	Fuzzy Attribute		
	High	Middle	Low
T1=5000	0.21	0.29	0.50
T2=15000	0.41	0.41	0.18
T3=10000	0.26	0.48	0.26
T4=20000	0.52	0.41	0.07
T5=2000	0.08	0.17	0.75

$Y' = \{High, Middle\}$, fuzzy support rate is

$$sup(Y') = \frac{0.21*0.41+0.41*0.48+\ldots+0.52*0.17}{5}$$

Temporal Fuzzy Association Rule

Example - Temporal

$$
\begin{array}{c c}
 & \begin{array}{ccccc} y_1 & y_2 & y_3 & \cdots & y_m \end{array} \\
\begin{array}{c} t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_n \end{array} &
\begin{pmatrix}
u_{11} & u_{12} & u_{13} & \cdots & u_{1m} \\
u_{21} & u_{22} & u_{23} & \cdots & u_{2m} \\
u_{31} & u_{32} & u_{33} & \cdots & u_{3m} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
u_{n1} & u_{n2} & u_{n3} & \cdots & u_{nm}
\end{pmatrix}
\end{array}
$$

$\mathrm{Y}' = \{High, Middle\}$, fuzzy support rate is

$$sup(Y') = \frac{0.21*0.41+0.41*0.48+\ldots+0.52*0.17}{5}$$

Discontinuous Temporal Fuzzy Item Set

Discontinuous temporal fuzzy item set $DY = \{Y_1 \xrightarrow{T_1} Y_2 \xrightarrow{T_2} Y_3 \xrightarrow{T_3} \cdots \xrightarrow{T_{c-1}} Y_c\}$, $T_i \neq 0, i \in \{1, \cdots, c-1\}$, where

$$Y_1 = \{x_1^1, x_2^1, \cdots, x_{p_1}^1\},$$
$$Y_2 = \{x_1^2, x_2^2, \cdots, x_{p_2}^2\},$$
$$\cdots,$$
$$Y_c = \{x_1^c, x_2^c, \cdots, x_{p_c}^c\},$$

are all consecutive item sets, $x_i^j \in Y = \{y_1, y_2, \cdots, y_m\}$.

Definition 5: For discontinues temporal fuzzy item set $DY = \{Y_1 \xrightarrow{T_1} Y_2 \xrightarrow{T_2} Y_3 \xrightarrow{T_3} \cdots \xrightarrow{T_{c-1}} Y_c\}$, the fuzzy support rate of DY is defined as:

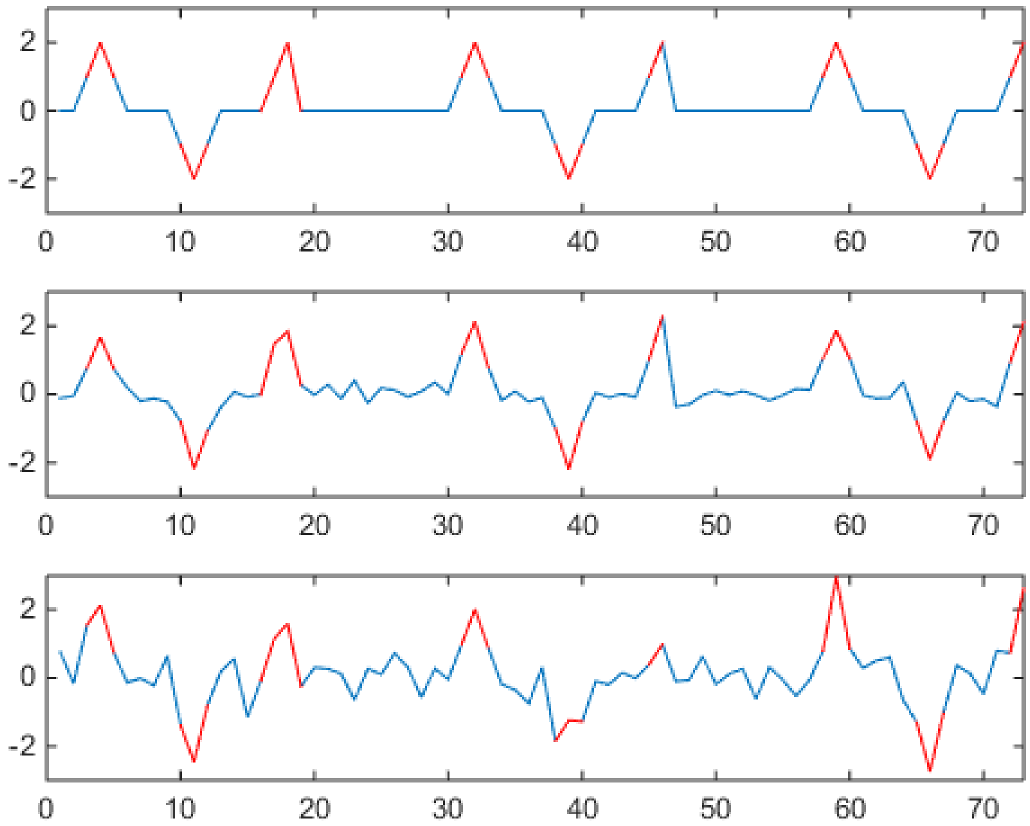
$$sup(DY) = \frac{\sum_{k=0}^{n-(p_1+p_2+\cdots p_c)} \max_{0 \le t_i \le T_i, 1 \le i \le c-1} \left(\prod_{i=1}^{c} \prod_{j=1}^{p_i} u\right)}{n - (p_1 + p_2 + \cdots + p_c)},$$

$$u = t_{k+(t_1+p_1)+\cdots+(t_{i-1}+p_{i-1})+j}\left(x_j^i\right).$$

where $t_i\left(x_j^i\right) = u_{ij}$ is the membership degree of granule t_i to cluster $x_i^j \in Y = \{y_1, y_2, \cdots, y_m\}$.

Discontinuous Temporal Fuzzy Item Set

Example – Discontinuous Rules



Complexity

Compute support rate in such definition is a Dynamic Programming.

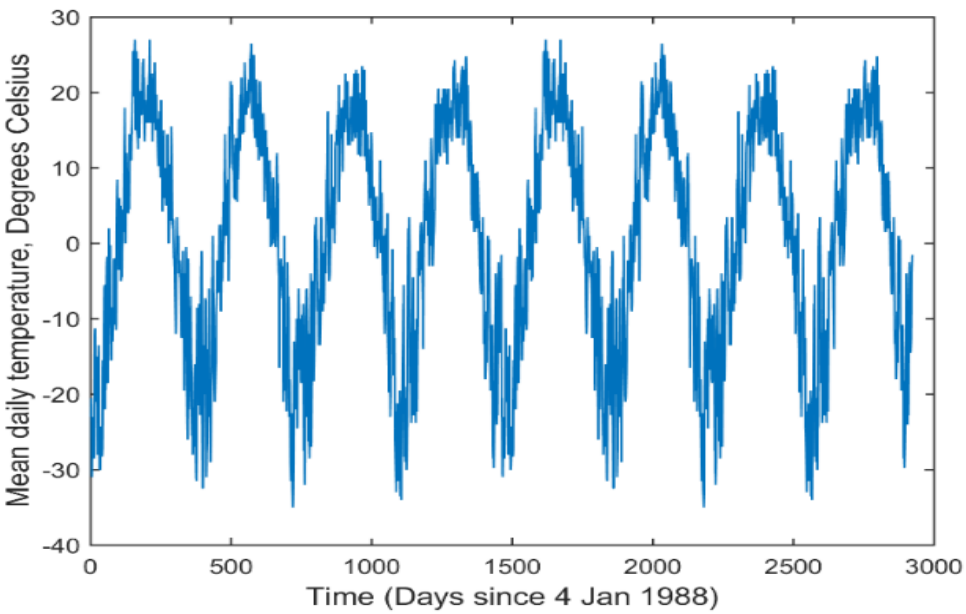
a) So that for a given $Y' = \{y_{i_1}, y_{i_2}, \cdots, y_{i_k}\}$ and position p, **we only need do 1 multiply operation** to figure out $Fsup[y_{i_1} y_{i_2} \cdots y_{i_k}][p]$.

b) Accordingly, for a given $Y = \{y_{i_1}, y_{i_2}, \cdots, y_{i_k}\}$, **we only need do $\boldsymbol{n - i_k \approx n}$ multiply operations** to figure out $sup\left(y_{i_1} y_{i_2} \cdots y_{i_k}\right)$.

c) Then, for discontinues item set $DY = \{Y_1 \xrightarrow{T_1} Y_2 \xrightarrow{T_2} Y_3 \xrightarrow{T_3} \cdots \xrightarrow{T_{c-1}} Y_c\}$, **we only need do $\boldsymbol{n \cdot T_1 \cdot T_2 \cdots T_{c-1}}$ multiply operations** to figure out $sup(DY)$.

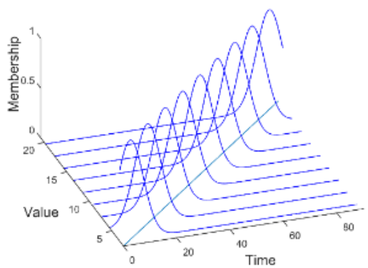
Experimental Study

The Mean Daily Temperature,
Fisher River near Dallas, Jan 01, 1988 to Dec 31, 1991

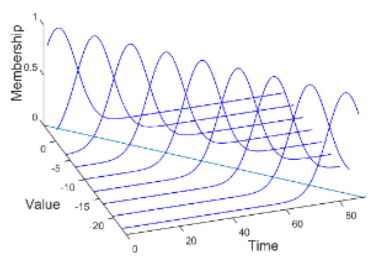


Experimental Study

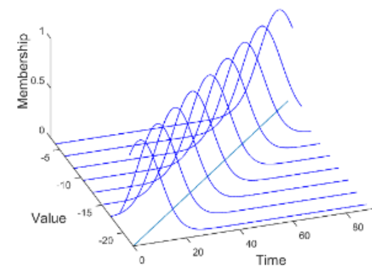
Four Clusters Based on LFIG



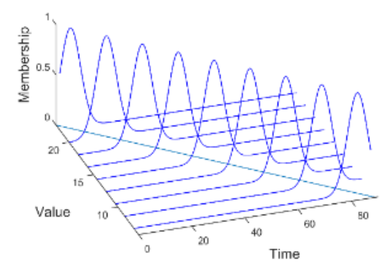
(a) the 1-st center *A1*



(b) the 2-nd center *A2*



(c) the 3-rd center *A3*



(d) the 4-th center *A4*

Experimental Study

Rules with Consecutive Sets

Rule	fuzzy association rules			
	antecedent	consequent	Support	Confidence
1	$A_4A_2A_3A_1$	A_4	0.1583	0.9796
2	$A_2A_3A_1$	A_4	0.1682	0.9792
3	A_3A_1	A_4	0.2059	0.9790
4	$A_4A_2A_3$	A_1	0.1659	0.9790
5	$A_1A_4A_2A_3$	A_1	0.1591	0.9760
6	A_2A_3	A_1	0.2215	0.9759
7	A_3A_1	A_1A_4	0.2213	0.9750
8	$A_1A_4A_2$	A_3	0.2036	0.9657

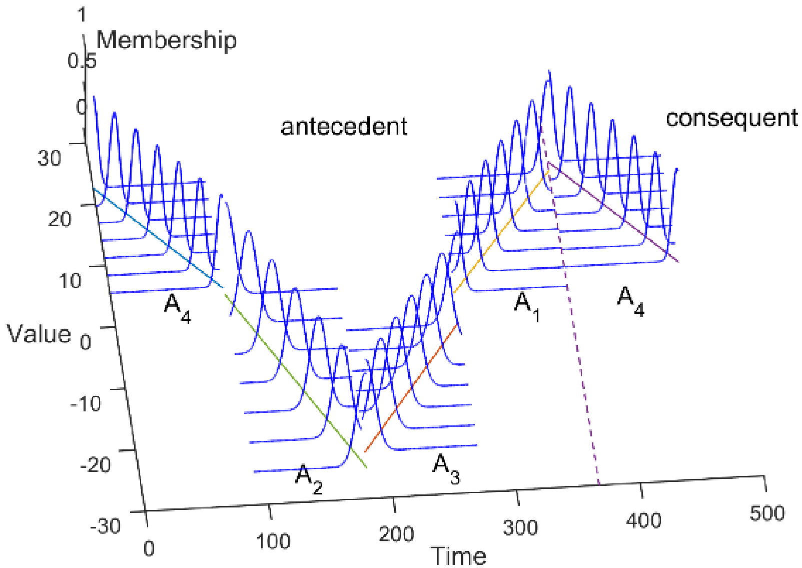
Experimental Study

Rules with Discontinues Sets

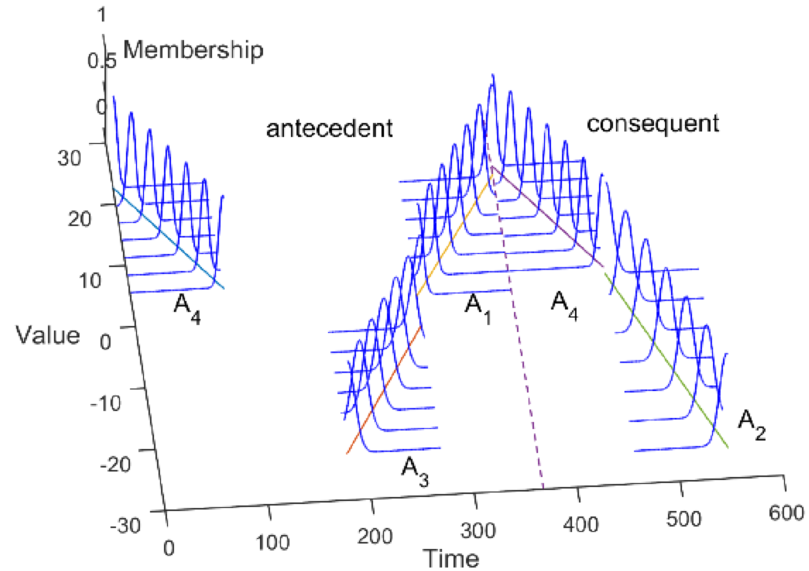
Rule	fuzzy association rules			
	antecedent	consequent	Support	Confidence
1	$A_1 \xrightarrow{T\leq 276} A_3$	A_1A_4	0.2135	0.9656
2	$A_1 \xrightarrow{T\leq 184} A_4$	A_2	0.2319	0.9154
3	$A_1 \xrightarrow{T\leq 184} A_2$	A_3	0.1987	0.9471
4	$A_3 \xrightarrow{T\leq 184} A_4$	A_2	0.2285	0.9540
5	$A_3 \xrightarrow{T\leq 184} A_4$	A_2A_3	0.2275	0.9530
6	$A_4 \xrightarrow{T\leq 184} A_3$	A_1	0.2136	0.9854
7	$A_4 \xrightarrow{T\leq 184} A_3A_1$	A_4A_2	0.1977	0.9357
8	$A_2 \xrightarrow{T\leq 184} A_1$	A_4A_2	0.2163	0.9983

Experimental Study

Visualization of Association Rules



Rule $A4\text{-}A2\text{-}A3\text{-}A1 \overset{T}{\Rightarrow} A4$



Rule $A4\text{-}A3\text{-}A1 \overset{T}{\Rightarrow} A4\text{-}A2$

Thanks

zebang@mail.bnu.edu.cn

July 28, 2017