

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

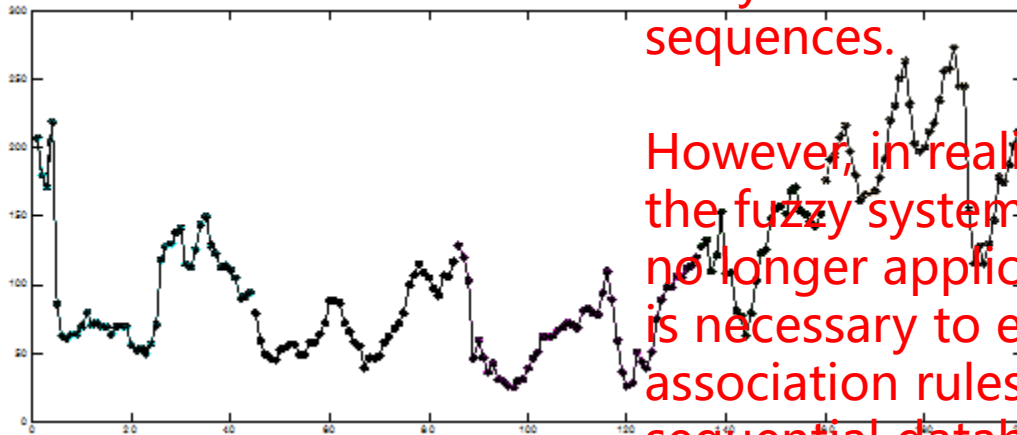
Contents

In this paper, we developed a sound conceptual framework for temporal fuzzy association rules mining based on fuzzy information granulation. I will mainly introduce our method by focusing general idea rather than formula and details.

- Introduction
- Prerequisites
- Temporal Fuzzy Association Rule
- Experimental Study
- Conclusion

Introduction

Work Flow of Our Work



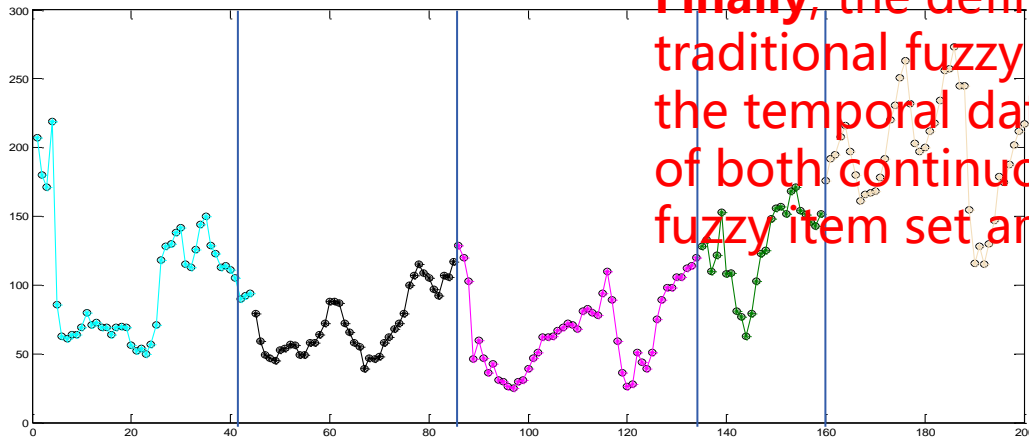
First, as we know, time association rules from time series in the fuzzy system is a challenging problem both in theory and in practice. Specifically, association rule learning typically does not consider the order of items either within a transaction or across transactions. In most studies, fuzzy association rules are not focused on time sequences.

However, in reality, there are many time series in the fuzzy system. Traditional association rules are no longer applicable to temporal data. Therefore, it is necessary to extend the definition of fuzzy association rules from traditional databases to time sequential databases. Such definitions should be reliable and appropriate to the rule mining process.

Introduction

This is our work flow. **First** we introduce some basic methods of fuzzy information granulation. In this paper, we applied a method called LFIG, published recently. **Second**, based on FCM clustering, partition matrix U is obtained. where u_{ij} is the membership degree of granule t_i to cluster y_j . **Finally**, the definition of the support rate of traditional fuzzy association rules is extended to the temporal data, including the fuzzy support rate of both continuous and discontinues temporal fuzzy item set and their association rules.

Work Flow of Our Work



$$\begin{matrix}
 & y_1 & y_2 & y_3 & \dots & y_m \\
 t_1 & u_{11} & u_{12} & u_{13} & \dots & u_{1m} \\
 t_2 & u_{21} & u_{22} & u_{23} & \dots & u_{2m} \\
 \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\
 t_N & u_{n1} & u_{n2} & u_{n3} & \dots & u_{nm}
 \end{matrix}$$

Fuzzy Information Granulation

Temporal Fuzzy Association Rule Mining

Fuzzy C-means for granular series



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.

Now Let's review some basic prerequisites.
Traditional association rule is A to B, which means if A occurs then B will occur.

Temporal association rule says, if A occurs then within time range of T, B will occur.

When A and B are fuzzy items, it is called fuzzy association rule.

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

$$f(x; kt + b, \sigma) = \exp\left(-\frac{(x - (kt + b))^2}{2\sigma^2}\right), \quad t \in [0, 1],$$

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

Fuzzy information granulation constitutes an important tool to provide appropriate solutions in predicting long-term future values, especially in fuzzy association rules mining. The time series is first broken down into successive pieces of simpler subseries and each subseries is then represented by a fuzzy set, referred to as fuzzy information granule. Consequently, the dimensionality of the problem and the computation overhead become greatly reduced. For example, Gaussian FIG is very common and remarkably useful among all fuzzy granules.

Recently, a latest published fuzzy information granule called LFIG, can be appropriate to represent the linear time trend by setting the core of a Gaussian fuzzy number μ to be linearly time-dependent. In our Paper, this method is applied.

Prerequisites

Fuzzy C-Means for Granular Time Series

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

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

The fuzzy partition matrix is U ,

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

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

In determining the structure in data, fuzzy clustering offers an important insight into data by producing gradual degrees of membership to individual patterns within clusters. A significant number of fuzzy clustering algorithms have been developed with widely known methods such as FCM. A finite collection of N patterns is described as $T = \{t_1, t_2, t_3, \dots, t_N\}$ (such pattern is a fuzzy granule in this paper) and collection of m cluster centers is denoted $Y = \{y_1, y_2, \dots, y_m\}$. The fuzzy partition matrix is U , where $u_{ij} = t_i(y_j)$ is the membership degree of granule t_i to cluster y_j , $i \in \{1, 2, \dots, N\}$, $j \in \{1, 2, \dots, m\}$.

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

Traditional fuzzy association is mostly based on transactional database or quantitative database with no sequential. For example, for the database of salaries, considered three classifications (fuzzy clusters) of “High”, “Middle”, “Low”, each salary has different membership on these three classifications. For example, S1 belongs to High by 0.21. (剩下照念此页右侧黑色文字即可)

$T = \{S1, S2, S3, S4, S5\}$ is a transaction set of salary.

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

For any sub set of Y , $Y' = \{y_1, y_2, \dots, y_p\}$, $y_i \in Y$, 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 $Y_1 \rightarrow Y_2$ is,

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

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

Prerequisites

照念即可 +

As you can see, if S1 becomes T1, which means salary has time order. For example, T1 means you get \$5000 at first month, T2 means you get \$15000 at second month and so on.

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.

Temporal Fuzzy Association Rule

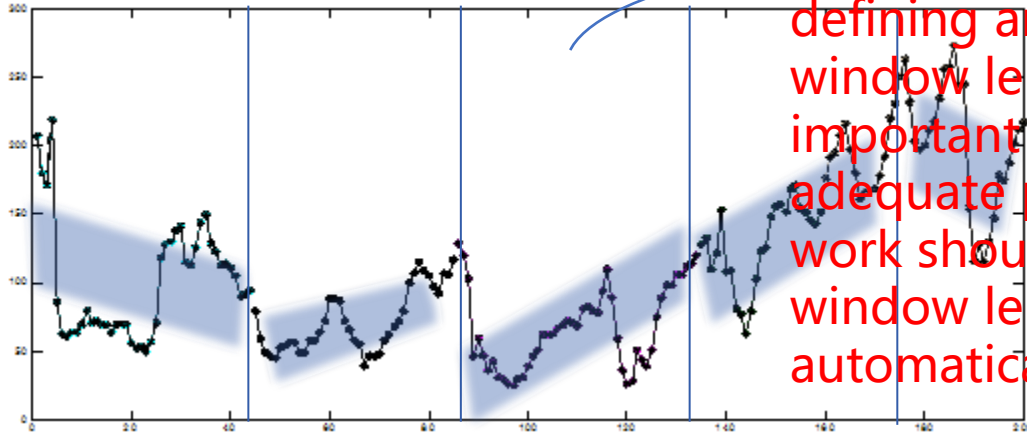
In this section

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

此页照念+However, an issue on how to find a fuzzy information graduation method matched with our fuzzy association rules mining when dealing with real world problems could be a subject of future studies. Especially, defining and selecting reliable time window lengths can be another important problem to provide adequate prediction accuracy. Future work should focus on how to select the window length effectively and automatically.



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

Temporal Fuzzy Association Rule

此页照念即可

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

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

Temporal Fuzzy Association Rule

此页照念+we will show you a simple example and everything of this definition.

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

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

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

First, let's see how traditional definition works. + 照念

$T = \{S_1, S_2, S_3, S_4, S_5\}$ is non-sequential dataset.

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


Fuzzy support rate of $Y' = \{y_1, y_2, \dots, y_p\}$, $y_i \in Y$, 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



照念+as you can see in the table, for the definition of support rate in traditional fuzzy association rules multiplications of memberships are done in each row of partition matrix U

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

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

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.

Now, for time series, in this simple example, fuzzy pattern is.....+照念

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

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

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. For example +照念 (注意箭头的动画效果)

Temporal Fuzzy Association Rule

Example - Temporal

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

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

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

In partition matrix U , it looks like this

Discontinuous Temporal Fuzzy Item Set

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

$$Y_1 = \{x_1^1, x_2^1, \dots, x_{p_1}^1\},$$

$$Y_2 = \{x_1^2, x_2^2, \dots, x_{p_2}^2\},$$

$\dots,$

$$Y_c = \{x_1^c, x_2^c, \dots, x_{p_c}^c\}$$

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

Now we can discuss discontinuous temporal fuzzy item sets. Notice that discontinuous sets are joined together with consecutive sets by time interval T . + 照念+we know that definition is a little bit complex, but in fact it's very simple. Let's see some examples.

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

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

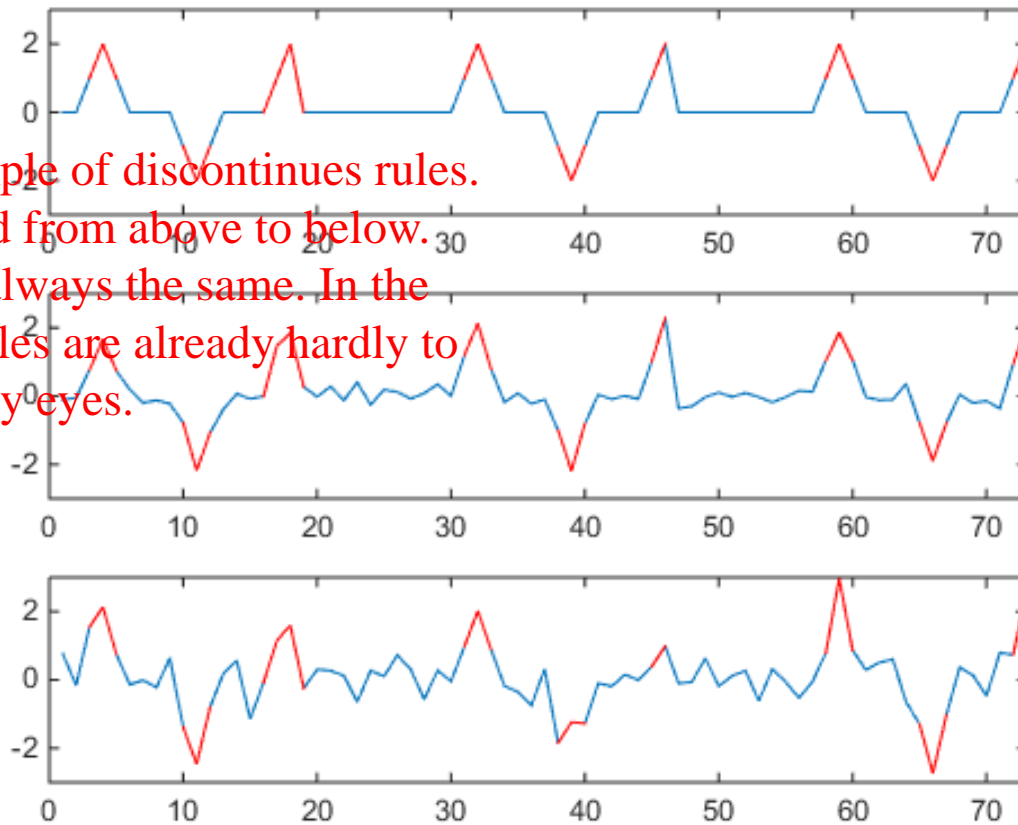
$$u = t_{k+(t_1+p_1)+\dots+(t_{i-1}+p_{i-1})+j}(x_j^i).$$

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

Discontinuous Temporal Fuzzy Item Set

Example – Discontinuous Rules

This is an example of discontinuous rules. noises increased from above to below. But the rule is always the same. In the third picture, rules are already hardly to be found only by eyes.



Complexity

As I said, computing support rate by our definition is very simple, which means the complexity is very low. As a matter of fact, + 照念以下

Compute support rate in such definition is a Dynamic Programming.

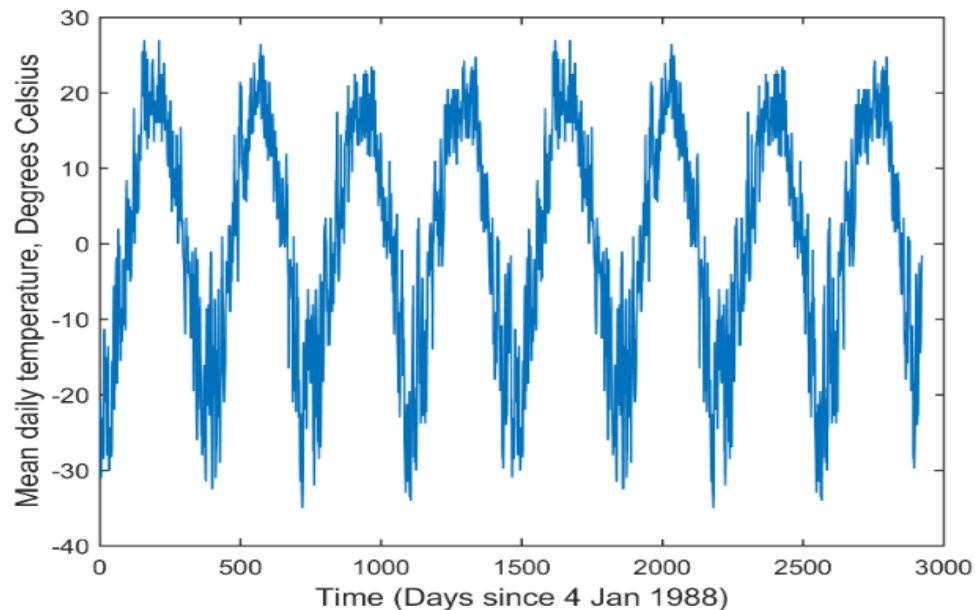
- a) So that for a given $Y' = \{y_{i_1}, y_{i_2}, \dots, y_{i_k}\}$ and position p , **we only need do 1 multiply operation** to figure out $Fsup[y_{i_1} y_{i_2} \dots y_{i_k}][p]$.
- b) Accordingly, for a given $Y = \{y_{i_1}, y_{i_2}, \dots, y_{i_k}\}$, **we only need do $n - i_k \approx n$ multiply operations** to figure out $sup(y_{i_1} y_{i_2} \dots y_{i_k})$.
- c) Then, for discontinues item set $DY = \{Y_1 \xrightarrow{T_1} Y_2 \xrightarrow{T_2} Y_3 \xrightarrow{T_3} \dots \xrightarrow{T_{c-1}} Y_c\}$, **we only need do $n \cdot T_1 \cdot T_2 \dots T_{c-1}$ multiply operations** to figure out $sup(DY)$.

Experimental Study

The Mean Daily Temperature,

Fisher River near Dallas, Jan 01, 1988 to Dec 31, 1991

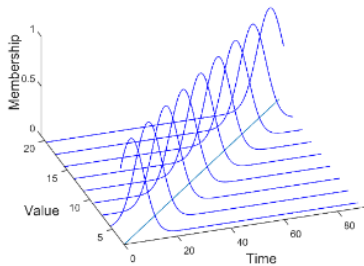
we briefly carried out an experiment by our definition and mining process on real time series. The mean daily temperature, is used for this experiment. We choose it because the length of window can be easily determined in such date. For this data, we expect to dig out the cyclical trend by our method.



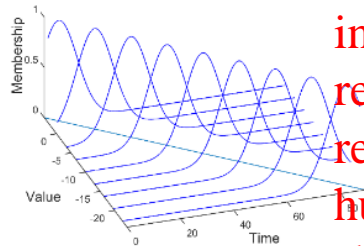
Experimental Study

Four Clusters Based on LFIG

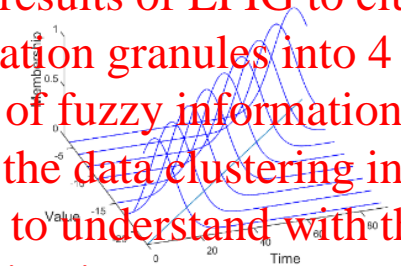
First, we set the length of a temporal window to be $T=92$ days, about a quarter of circle and such length does not cut off the cycle of periodic data, then the whole time series is divided into 33 temporal windows. Now we translate the original time series into the time window series. For each time window, we apply LFIG method to generate the fuzzy information granules. Second, we apply FCM on the results of LFIG to cluster the 33 fuzzy information granules into 4 classifications. The results of fuzzy information granulation can reflect the data clustering in a way easy for human to understand with the aids of LFIG, as shown in picture.



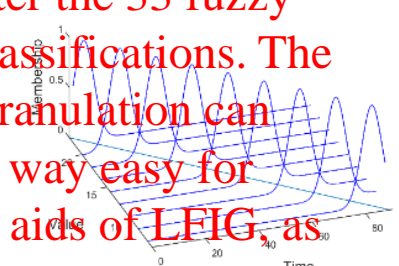
(a) the 1-st center $A1$



(b) the 2-nd center $A2$



(c) the 3-rd center $A3$



(d) the 4-th center $A4$

Experimental Study

By fuzzy association rule mining discussed above, we set the threshold of support ≥ 0.158 and confidence ≥ 0.9657 , the forecasting length $T = 92$. Then we have the strong fuzzy association rules listed in the Table by order of confidence. Rule 1 means: if A_{N-3} ' is A_4 , A_{N-2} ' is A_2 , A_{N-1} ' is A_3 , A_N ' is A_1 , then A_{N+1} ' is A_4 .

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

At last, we also found the association rules that are generated by discontinuous item sets.

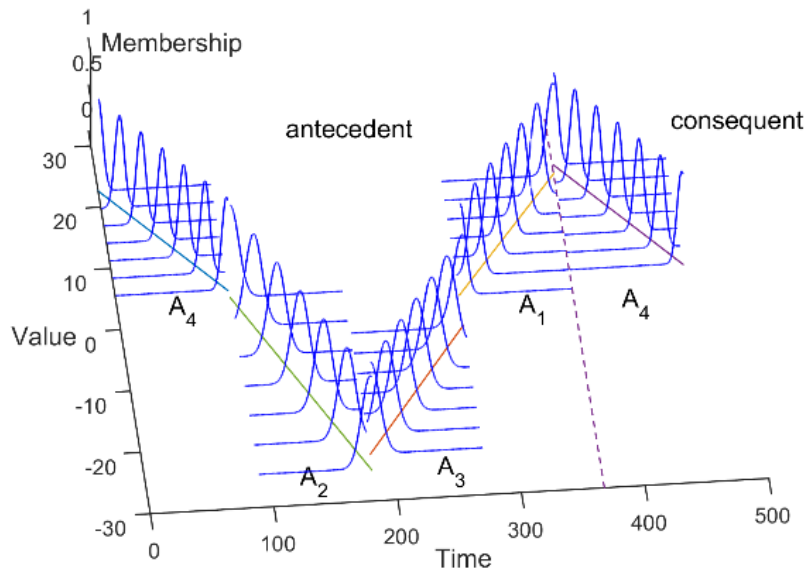
Rule 1 means: if A_{N-3} ' is A_1 , A_N ' is A_3 , then A_{N+1} ' A_{N+2} ' is A_1A_4 .

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

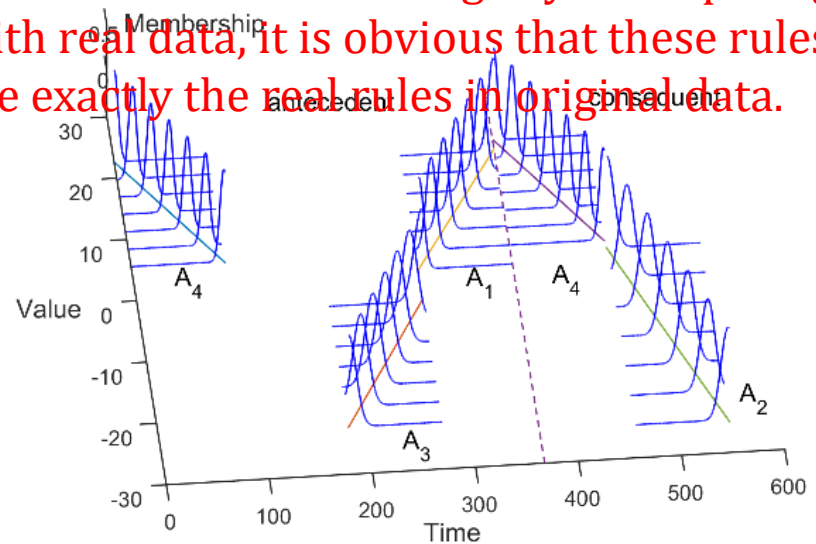
Experimental Study

Visualization of Association Rules



Rule $A4-A2-A3-A1 \xRightarrow{T} A4$

In order to see these rules more clearly and to illustrate the reliability of these rules, we visualize the first rule in both Table. For example, on the right side, Rule $A4-A3-A1 \xRightarrow{T} A4-A2$ means if $A4$ occurs and $A3A1$ occurs within 276 days, then $A4A2$ will occur in the following days. Comparing with real data, it is obvious that these rules are exactly the real rules in original data.



Rule $A4-A3-A1 \xRightarrow{T} A4-A2$



Thanks

zebang@mail.bnu.edu.cn

July 28, 2017