Temporal Fuzzy Association Rules Mining Based on Fuzzy Information Granulation

Zebang Li, Fan Bu, Fusheng Yu
School of Mathematical Sciences, Beijing Normal University
Laboratory of Mathematics and Complex Systems, Ministry of Education
Beijing, People's Republic of China

Abstract—**In this paper, we developed a sound conceptual framework for temporal fuzzy association rules mining based on fuzzy information granulation. First, 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. It is found that the computation of support rate under our definition is a simply dynamic programming problem with very low complexity. Then an algorithm based on Apriori is demonstrated. At last, experiments were carried out to show the good performance of our new algorithm in mining continuous rules and discontinuous rules by our definition.**

Keywords-time series; temporal fuzzy association rules; fuzzy information granulation; fuzzy clustering

I. Introduction

Time association rules from time series in the fuzzy system is a challenging problem both in theory and in practice, but association rule learning typically does not consider the order of items either within a transaction or across transactions.

Association rule mining forms an important research area in the field of data mining. Fuzzy association rules solve the problem of sharp boundaries [1][2]. Association rule mining is a process of finding association rules in a dataset [3]. Association rule is of the form A→B where A and B are disjoint subsets having minimum user specified support and confidence level.

Association rules can be divided based on item sets of interest as positive and negative [4]. Another classification of association rules is Boolean, quantitative and fuzzy association rules [4].

Fuzzy association rule mining is mostly based on Apriori [5], SLP-growth algorithm [6] and FP-growth algorithm [5]. These methods have been mostly used in data mining as a model for prediction a target value based on a given relational database.

In most studies, fuzzy association rules are not focused on time sequences [7]. 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 Traditional association rules are no longer applicable to temporal data.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.

In this paper, we develop a sound conceptual framework for temporal fuzzy association rules mining based on fuzzy information granulation. Section II demonstrates some prerequisites. In section III, we extend the definition of support rate of fuzzy association rules from non-sequential data to temporal data and we showed that the new proposed mining algorithm based on our new definition has a very low computation complexity. In section IV, experiments were carried out to show the good performance of our new algorithm in mining continues and discontinues rules by our definition.

II. Prerequisites

In this section, we introduced the concept of traditional and a new type of fuzzy information granule. Also, the concept of fuzzy c-means and fuzzy association rules based on non-sequential series are presented.

A. Fuzzy Information Granule

Fuzzy information granulation constitutes an important tool to provide appropriate solutions in predicting long-term future values, especially in fuzzy association rules mining [8]. 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 (FIG) [9][10]. Consequently, the dimensionality of the problem and the computation overhead become greatly reduced.

Fuzzy information granules include triangular FIG, interval FIG, and Gaussian FIG [11][12]. For example, Gaussian FIG is very common and remarkably useful among all fuzzy granules. A Gaussian membership function f of a granule is specified by two parameters μ and σ as follows:

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

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

However, all of these methods of FIG do not represent the time trend, which lost the key information and prevent future prediction.

B. Linear Gaussian Fuzzy Information Granule and Its Distance Measure

Recently, a new type of FIG is presented called Linear Gaussian FIG by [13]. Traditional FIG only represents the meaning without time trend such as "small", "large". Instead, Linear Gaussian FIG, we called LFIG, is the modification of traditional Gaussian FIG. 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. It is proposed to overcome disadvantages of the traditional information granulation. Distance measurement along with this new type of granules is also presented in [13].

Definition 1: A linear Gaussian fuzzy number $LG(k, b, \sigma, T)$ has a membership function such that at a given time $t \in [0, T]$, the membership grade of a value x belonging to G is expressed as:

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

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.

If $\mu(t)$ is regarded as some constant, then $LG(k, b, \sigma, T)$ degenerates into $G(\mu, \sigma)$. $\mu(t) = kt + b$ reflects a linear trend of change in the time window. Parameters of the LFIG, namely, the core line regression line $\mu(t) = kt + b$ and the spread σ, can be easily determined by linear regression [14]. Linear regression is intuitively appealing, not only because its results are components of information granules, but it comes with an acceptable computation complexity. Give a subseries $Y = \{Y_{T_0+1}, Y_{T_0+2}, \cdots, Y_{T_0+T}\}$ coming from the original time series, we treat Y_{T_0+1} as the start point of Y, and rewrite Y in the form $Y = \{Y_1, Y_2, \cdots, Y_T\}$. A linear regression analysis

$$Y_t = kt + b + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$, is then applied to generate a LFIG $LG(k, b, \sigma, T)$. The trend of change and deviation range in the time window can be represented by the core $\mu(t) = kt + b$, and σ^2. A linear Gaussian fuzzy number $LG(k, b, \sigma, T)$ can be demonstrated as Fig. 1.

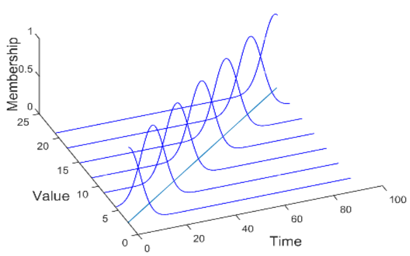


Fig. 1 linear fuzzy information granule

C. Fuzzy C-Means (FCM)

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 [15]. A finite collection of N patterns is described as $T = \{t_1, t_2, t_3, \cdots, t_N\}$ (such pattern is a fuzzy granule in this paper) and collection of m cluster centers is denoted $Y = \{y_1, y_2, \cdots, y_m\}$. 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, $i \in \{1,2, \cdots, N\}$, $j \in \{1,2, \cdots, m\}$.

D. Traditional Fuzzy Association Rule

Traditional fuzzy association is mostly based on transactional database or quantitative database with no sequential. For example, for the database of salaries [16], considered three classifications (fuzzy clusters) of "High", "Middle", "Low", each salary has different membership on these three classifications (Table I cited from [16]).

TABLE I. 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 [16] 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'. In this case, $n = 5$ and $p = 3$. However, such definition does not consider the time trend. When the database$T = \{x_1, x_2, x_3, \cdots, x_N\}$ is time series, new definition of support rate needs to be presented.

Fuzzy association rules such as $Y_1 \to Y_2$ represents the case: if Y_1 occurs, then Y_2 will occur. Y_1 is called antecedent of the rule and Y_2 is called consequent of the rule. p, q are the length of Y_1, Y_2, then fuzzy support rate of association rule $Y_1 \to Y_2$ is,

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

III. TEMPORAL FUZZY ASSOCIATION RULE BASED ON FUZZY INFORMATION GRANULATION

In this section, we extended the definition of fuzzy association rule from none-sequential to sequential. Such promotion is based on the traditional definition of fuzzy association rules and the partition matrix, which is the result of fuzzy information granulation and fuzzy clustering. At last, we showed that our definition has a very low computation complexity.

A. The Generalization of Fuzzy Association Rules Support Definition in Time Series

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, where l is window size. Granular time series $T' = \{t_1, t_2, t_3, \cdots, t_{N-l}\}$ is clustered by fuzzy c-means [11]. $Y = \{y_1, y_2, \cdots, y_m\}$ is the clusters of fuzzy c-means, which has m classifications.

Based on FCM clustering, partition matrix U is obtained. The structure of partition matrix is:

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

where $u_{ij} = t_i(y_j)$ is the membership degree of granule t_i to cluster y_j. And for time series, each row of partition matrix U is not independent any more. The order of rows in matrix U represents the order of time.

Definition 2: Given that Y is a set of fuzzy clusters and $y_{i_j} \in Y$ is the cluster such as "high" or "low", consecutive temporal fuzzy item set is defined as $Y' = \{y_{i_1}, y_{i_2}, \cdots, y_{i_p}\}$, $y_{i_j} \in \{y_1, y_2, \cdots, y_m\}$, which $y_{i_i}, y_{i_2}, \cdots, y_{i_p}$ occurs followed by continues time order and it may be repeated in Y_1 and Y_2. The adjacent time interval in a consecutive temporal fuzzy item set is zero.

For example, $Y_1 = \{y_1, y_1, y_1\}$ represents the first day is y_1, second and third day is still y_1.

Association rule learning typically does not consider the order of clusters either within a transaction or across transactions. That is to say, for the definition of support rate in traditional fuzzy association rules multiplications of memberships are done in each row of partition matrix U, as introduced in section II. Now, 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, y_1 y_2 y_3 occurs by successive time order. Then the $u_{11} \cdot u_{22} \cdot u_{33}$ presents the fuzzy membership of the case when t_1 is y_1, t_2 is y_2, t_3 is y_3. Intuitively speaking, the definition of support rate for traditional item sets is to multiply membership degrees in each row of partial matrix U and then sum the results. When it comes to sequential data, multiplication must be across each row of matrix U.

Consider the case of consecutive temporal fuzzy item set first and then the situation of discontinues can be discussed. $Y' = \{y_{i_1}, y_{i_2}, \cdots, y_{i_p}\}$ is a consecutive temporal fuzzy item set. $y_1, y_2, \cdots, y_p$ occur followed by continues time order.

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. $\sum_{k=0}^{n-p} \prod_{j=1}^{j=p} u_{j+k,i_j}$ is called fuzzy support number of consecutive temporal fuzzy item set Y', written as $Fsup(Y')$. Frequent temporal fuzzy item sets are those whose $Sup(Y') \geq \mu$, where μ is the threshold of frequent temporal fuzzy item set.

Example 3.1: For $Y' = \{y_1, y_2, y_3\}$, to compute the fuzzy support rate $Sup(Y')$, we first examine the fuzzy support number of the case when the t_1 is y_1, t_2 is y_2, t_3 is y_3, which is $t_1(y_1) \cdot t_2(y_2) \cdot t_3(y_3) = u_{11} \cdot u_{22} \cdot u_{33}$. Then when we slide the time window to the next to examine the fuzzy support number of the case when the t_2 is y_1, t_3 is y_2, t_4 is y_3, which is $t_2(y_1) \cdot t_3(y_2) \cdot t_4(y_3) = u_{21} \cdot u_{32} \cdot u_{43}$, and so on. The total fuzzy support number is

$$Fsup(Y) = t_1(y_1) \cdot t_2(y_2) \cdot t_3(y_3) + t_2(y_1) \cdot t_3(y_2) \cdot t_4(y_3) + \cdots + t_{n-2}(y_1) \cdot t_{n-1}(y_2) \cdot t_n(y_3).$$

So that,

$$sup(Y) = \frac{Fsup(Y)}{n-3} = \frac{\sum_{k=0}^{n-3} \prod_{j=1}^{j=3} u_{j+k,i_j}}{n-3}.$$

Particularly, when $Y' = \{y_{i_1}\}$, $Fsup(Y')$ is just the sum of partition matrix U by columns.

Next, consider the definition of the fuzzy association rules about consecutive temporal fuzzy item set. In our discussion, let $Y = \{y_1, y_2, \cdots, y_m\}$ be the fuzzy clusters. $Y_1 = \{y_{i_1}, y_{i_2}, \cdots, y_{i_p}\}$ and $Y_2 = \{y_{i_{p+1}}, y_{i_{p+2}}, \cdots, y_{i_{p+q}}\}$ are subsets of Y, noticed that $y_{i_j} \in Y$ is the fuzzy cluster such as "high" or "low", and it may be repeated in Y_1 and Y_2.

The temporal association rules such as $Y_1 \overset{T}{\Rightarrow} Y_2$ represents the case: if Y_1 occurs, then Y_2 will occur within the time range of T. Y_1 is called antecedent of the rule and Y_2 is called consequent of the rule. For fuzzy support rate of temporal association $Y_1 \overset{T}{\Rightarrow} Y_2$, we first examine the fuzzy support number of antecedent Y_1 based on Definition 3, then examine the consequent Y_2 within the following T time granules.

Definition 4: For consecutive temporal fuzzy item set $Y_1 = \{y_{i_1}, y_{i_2}, \cdots, y_{i_p}\}$ and $Y_2 = \{y_{i_{p+1}}, y_{i_{p+2}}, \cdots, y_{i_{p+q}}\}$, the fuzzy support rate of the fuzzy association rules $Y_1 \overset{T}{\Rightarrow} Y_2$ is defined as:

$$sup\left(Y_1 \overset{T}{\Rightarrow} Y_2\right) = \frac{\sum_{k=0}^{n-p} \max\limits_{0 \leq t \leq T} \left(\prod_{j=1}^{j=p+q} u_{j+k+t,i_j}\right)}{n-p}.$$

Accordingly, the fuzzy confidence rate is defined as:

$$conf\left(Y_1 \overset{T}{\Rightarrow} Y_2\right) = \frac{sup\left(Y_1 \overset{T}{\Rightarrow} Y_2\right)}{sup(Y_1)}.$$

Notice that in our definition, association rule $Y_1 \overset{T}{\Rightarrow} Y_2$ means Y_2 occurs within T after Y_1, so we use the function *max* to present the meaning of Y_2 occurring after Y_1 within the time range T, which means that once Y_2 occurs in the time range T after Y_2, then the association rule comes into existence.

Now we can discuss discontinuous temporal fuzzy item sets. Notice that discontinuous sets are joined together with consecutive sets by time interval T. For example, discontinuous temporal fuzzy item set $DY = \{Y_1 \overset{T_1}{\rightarrow} Y_2 \overset{T_2}{\rightarrow} Y_3 \overset{T_3}{\rightarrow} \cdots \overset{T_{c-1}}{\longrightarrow} Y_c\}$, $T_i \neq 0, i \in \{1, \cdots, c-1\}$, 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, $x_i^j \in Y = \{y_1, y_2, \cdots, y_m\}$, $i \in \{1,2, \cdots, p_j\}$, $j \in \{1,2, \cdots, c\}$.

Similar to the computation of support rate of association rules in consecutive sets, the support rate of discontinues set $DY = \{Y_1 \overset{T_1}{\rightarrow} Y_2 \overset{T_2}{\rightarrow} Y_3 \overset{T_3}{\rightarrow} \cdots \overset{T_{c-1}}{\longrightarrow} Y_c\}$ can be determined by Definition 5 as below. Accordingly, the definition of support rate of fuzzy association rules generated by discontinues sets were also given by Definition 6.

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

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

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

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

Definition 6: For discontinues temporal fuzzy item set $DY = \{Y_1 \overset{T_1}{\rightarrow} Y_2 \overset{T_2}{\rightarrow} Y_3 \overset{T_3}{\rightarrow} \cdots \overset{T_{c-1}}{\longrightarrow} Y_c\}$ and the consequent $Y_{c+1} = \{x_1^{c+1}, x_2^{c+1}, \cdots, x_{p_{c+1}}^{c+1}\}$. The fuzzy support rate of $Rule = \{DY \overset{T_c}{\Rightarrow} Y_{c+1}\}$ is defined as:

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

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

where $t_i(x_j^i) = u_{ij}$ is the membership degree of granule t_i to cluster $x_i^j \in Y = \{y_1, y_2, \cdots, y_m\}$. As we can see, this Definition is a special case of Definition 5 when c replaced by $c + 1$.

Accordingly, the fuzzy confidence rate is expressed as:

$$conf\left(DY \overset{T_c}{\Rightarrow} y\right) = \frac{sup\left(DY \overset{T_c}{\Rightarrow} y\right)}{sup(DY)}.$$

B. *Temporal Fuzzy Association Rule Mining*

In this part, we first designed the algorithm of fuzzy association rule mining based on our definition and Apriori algorithm [4]. Second, we showed that our definition has a relatively low computational complexity.

1) *Algorithm of Rule Mining:*

Input: time series, length of intervals, minimum support and confidence.

Output: fuzzy association rules.

Step1. Translate the original time series into granular time series by LFIG. The partition matrix U is obtained by and FCM.

Step2. Find the consecutive frequent temporal fuzzy item set by Definition 3. Similar to Apriori, start with 1-frequent-set and extend them to larger and larger item sets as long as those item sets appear sufficiently often in the database.

Step3. Find the discontinues frequent temporal fuzzy item set by Definition 5. We put the consecutive frequent sets together. First find the $Y_1 \overset{T_1}{\rightarrow} Y_2$. Then extend them to $Y_1 \overset{T_1}{\rightarrow} Y_2 \overset{T_2}{\rightarrow} Y_3 \overset{T_3}{\rightarrow} \cdots$ until they do not meet the threshold.

Step4. Find fuzzy association rules by Definition 6. Candidate antecedents of rules are those frequent patterns obtained in step two and three.

2) *Low Computation Complexity of Support Rate:*

The most important part of the flow is how to compute the fuzzy support rate of item set Y'. The definition of fuzzy support rate of Y' is given by Definition 3 or Definition 5 based on whether Y' is consecutive or discontinues. Although the expressing of Definition 3 seems very complex, we do not need to compute each multiplication in the expressing. In fact, the computation of the support of consecutive temporal fuzzy item sets is a simple dynamic programming problem with very low complexity. For convenience of description, the following definitions are given first.

Definition 7: The fuzzy support number of consecutive item set $Y' = \{y_{i_1}, y_{i_2}, \cdots, y_{i_k}\}$ at time position p (y_{i_k} at t_p, $y_{i_{k-1}}$ at t_{p-1} and so on) is written as $Fsup[y_{i_1} y_{i_2} \cdots y_{i_k}][p]$.

Example 7.1: for $Y' = \{y_1, y_2, y_3\}$, $p = 3$, $Fsup[y_1 y_2 y_3][p]$ represents the fuzzy support number of the case when the t_1 is y_1, t_2 is y_2, t_3 is y_3.

For time series with the length of n, the relationship between $Fsup[y_{i_1}, y_{i_2}, \cdots, y_{i_k}][p]$ and $\sup(y_{i_1} y_{i_2} \cdots y_{i_k})$ is:

$$sup(y_{i_1} y_{i_2} \cdots y_{i_k}) = \frac{\sum_{p=i_k}^{n} Fsup[y_{i_1} y_{i_2} \cdots y_{i_k}][p]}{n - i_k}.$$

Obviously, we have the recurrence relation as:

$$Fsup[y_{i_1} y_{i_2} \cdots y_{i_k}][p] = Fsup[y_{i_1} y_{i_2} \cdots y_{i_{k-1}}][p-1] \cdot Fsup[y_{i_k}][p],$$

where $Fsup[y_{i_1} y_{i_2} \cdots y_{i_k}][p]$ is an unknown quantity while $Fsup[y_{i_1} y_{i_2} \cdots y_{i_{k-1}}]$ and $Fsup[y_{i_k}][p]$ are both already known.

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 $n - i_k \approx n$ multiply operations to figure out $sup(y_{i_1} y_{i_2} \cdots y_{i_k})$.

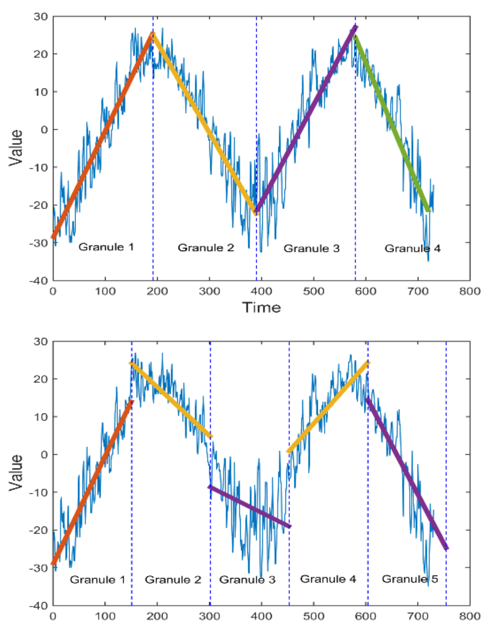
c) Then, for discontinues item set $DY = \{Y_1 \xrightarrow{T_1} Y_2 \xrightarrow{T_2} Y_3 \xrightarrow{T_3} \cdots \xrightarrow{T_{c-1}} Y_c\}$, we only need do $n \cdot T_1 \cdot T_2 \cdots T_{c-1}$ multiply operations to figure out $sup(DY)$.

IV. Experimental Study

In this section, we briefly carried out an experiment by our definition and mining process on real time series. The fuzzy information granule of LFIG [9] introduced in Section II are used in our experiment. Results showed good performance of the experiment.

A. The Length of Window of Information Granule

The length of granule window is an important parameter in the proposed algorithm. Final results depend on the size of the window. Either the window is too small or big, some windows may cut off the cycle of periodic data, then the results of fuzzy granulation cannot be good. Suitably chosen length can capture the varying pattern of a time series. However, an unsuitably chosen length may lead to poor data summarization and render the fuzzy inference system worthless. As shown in Fig. 2, when the length of time window is properly set to be 190, the folding core lines of the 4 generated LFIGs can precisely and concisely reflect the changes of the time series. However, when the length is set to be 150, the generated 5 LFIGs cut off the circle of the data, which means the time series deviate more serious from the core lines, thus these 5 LFIGs cannot represent the time series very well.



(a) A time series with 4 windows (b) A time series with 5 windows

Fig.2. time series with different window sizes

However, this problem is beyond the topic of this paper, so we choose the cycle data and try to set the window size to be just conform to the cycle of the data. In the following experiments, if a time series have a natural time cycle, the time window is often set to be a half or a quarter of the cycle. By such setting we hope the trends among a cycle can be detected and the relation among these trends are retained.

B. Experimental Study: mean daily temperature of Fisher River near Dallas

The mean daily temperature, Fisher River near Dallas, Jan 01, 1988 to Dec 31, 1991 [17] (Fig. 3) 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.

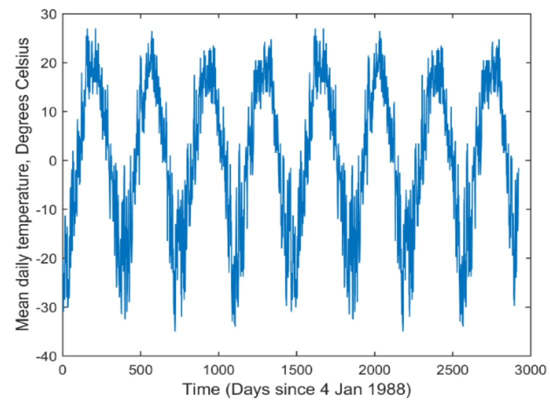


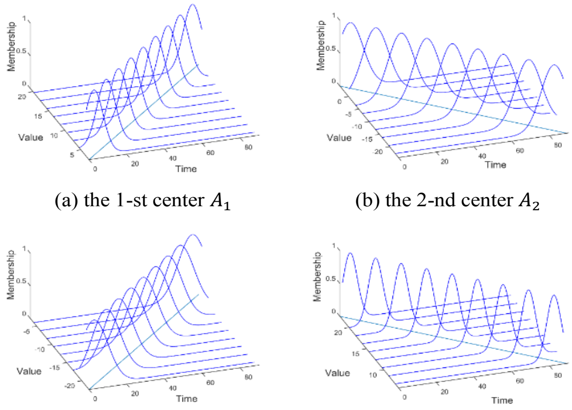
Fig. 3. The mean daily temperature, Fisher River near Dallas, Jan 01, 1988 to Dec 31, 1991

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 ($length(data) = 33 \times 92 = 2976$). Now we translate the original time series $data = \{t'_1, t'_2, \cdots, t'_n\}$ ($n = 2976$) into the time window series $data' = \{t_1, t_2, \cdots, t_N\}$ ($N = 33$). For each time window, we apply LFIG method to generate the fuzzy information granules $A_i = LG(k_i, b_i, \sigma_i, T), i = 1,2, \cdots, N$, which $N = 33$, $T = 92$. The definition of LFIG is introduced in section II.

Second, we apply FCM on the results of LFIG to cluster the 33 fuzzy information granules into 4 classifications. The measurement of distance of each granule is given by [4]. In detail, the four clusters are the center of all granules, which are also the fuzzy information granules $A_i = LG(k_i, b_i, \sigma_i, T), i = 1,2,3,4, T = 92$:

$$A_1 = (k_{90}, b_{90}, \sigma_{90}) = (0.2057, 2.4823, 4.8886),$$
$$A_2 = (k_{91}, b_{91}, \sigma_{91}) = (-0.1898, 22.6379, 3.7485),$$
$$A_3 = (k_{92}, b_{92}, \sigma_{92}) = (-0.3203, 4.1758, 6.8577),$$
$$A_4 = (k_{90}, b_{90}, \sigma_{90}) = (0.2191, -22.3964, 7.4308).$$

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 Fig. 4.



(a) the 1-st center A_1 (b) the 2-nd center A_2

(c) the 3-rd center A_3 (d) the 4-th center A_4

Fig. 4 four classifications of fuzzy information granules.

Then firstly, consider the association rules that are generated by the consecutive sets. By fuzzy association rule mining discussed above, we set the threshold of sup ≥ 0.158 and conf ≥ 0.9657, the forecasting length of $Y_1 \overset{T}{\Rightarrow} Y_2, T = 92$. Then we have the strong fuzzy association rules listed in the Table II by order of confidence.

TABLE II. STRONG FUZZY ASSOCIATION RULES[A]

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

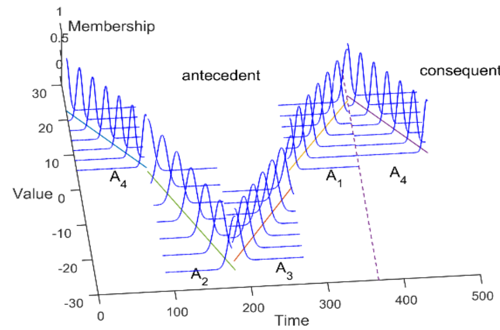
a. Support≥0.158 and Confidence≥0.9657

Rule 1: 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,

Rule 2: if A'_{N-2} is A_2, A'_{N-1} is A_3, A'_N is A_1, then A'_{N+1} is A_4,

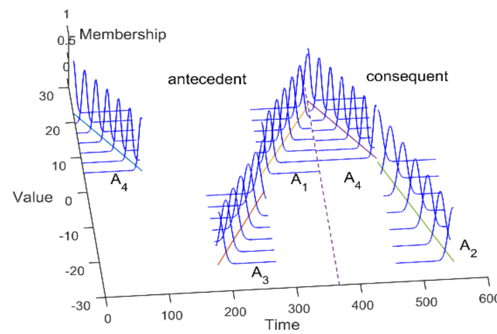
⋮

For example, to illustrate the reliability of these rules, we visualize the first rule in the Table II, which is $\{A_4 \rightarrow A_2 \rightarrow A_3 \rightarrow A_1\} \overset{T}{\Rightarrow} \{A_4\}$. The antecedent is $\{A_4 \rightarrow A_2 \rightarrow A_3 \rightarrow A_1\}$ and the consequent is $\{A_4\}$. It can be seen as Fig. 5, which means that if the pattern of $\{A_4 \rightarrow A_2 \rightarrow A_3 \rightarrow A_1\}$ occurs then A_4 will occur within the time range of T.



Rule $A_4A_2A_3A_1 \overset{T\leq 92}{\Longrightarrow} A_4$

Fig.5. Visualization of Association Rules.



Rule $A_4 - A_3A_1 \overset{T\leq 92}{\Longrightarrow} A_4A_2$

Fig.6. Visualization of Association Rules.

Comparing with real data, it is obvious that these rules are exactly the real rules in original data. In detail, as we can see, the rules in Table II can be interpreted as $A_4 \rightarrow A_2 \rightarrow A_3 \rightarrow A_1 \rightarrow A_4 \rightarrow A_2 \rightarrow A_3 \rightarrow A_1 \rightarrow \cdots$ and it is just the real circle of the original time series.

At last, we found the association rules that are generated by discontinuous item sets. We set the threshold of sup ≥ 0.197 and conf ≥ 0.91, the forecasting length of $Y_1 \overset{T}{\Rightarrow} Y_2, T = 92$. Then we have the strong fuzzy association rules listed in the Table III. There are so many discontinues rules and we only demonstrate a part of them.

TABLE III. FUZZY ASSOCIATION RULES WITH DISCONTINUES SETS[A]

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

Rule	fuzzy association rules			
	antecedent	*consequent*	*Support*	*Confidence*
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

a. they are only a part of the whole rules.

Rule 1: if A'_{N-3} is A_1, A'_N is A_3, then $A'_{N+1}A'_{N+2}$ is A_1A_4,

Rule 2: if A'_{N-2} is A_1, A'_N is A_4, then A'_{N+1} is A_2,

⋮

In the same way, to illustrate the reliability of these rules, we visualize Rule 7 as Fig. 6. Comparing with original time series, it is obvious that these rules are exactly the real rules in original data.

V. CONCLUTIONS

The idea of fuzzy association rules is essential to discover the interesting knowledge in data and fuzzy information granulation is effective in data mining. However, the traditional definition of the support rate of fuzzy association based on fuzzy information granulation cannot be applied to the time sequential data.

In this paper, we extend the definition of support rate of fuzzy association rules from non-sequential data to temporal data. Our definitions are not only reasonable in demonstrating the inherent meaning of fuzzy temporal association rules but also has a very low computation complexity. Because of our definitions of support rate of fuzzy association rules are only related to partition matrix produced by fuzzy information granulations and clustering. It is easy to be found that the computation of support rate can be a simply dynamic programming problem with very low complexity. Our method combines the fuzzy information granulation and fuzzy association rules mining. The generalization of support rate on time-dependent fuzzy association rules may improve the ability to discover the interesting knowledge in complicated time series.

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.

ACKNOWLEDGMENT

This work is supported by the National Training Programs of Innovation and Entrepreneurship, the National Natural Science Foundation of China (No.11571001), and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] MY Chen，BT Chen, “A hybrid fuzzy time series model based on granular computing for stock price forecasting,” Information Sciences, 294 (2015) 227-241.

[2] W. Wang, W. Pedrycz, X. Liu, “Time series long-term forecasting model based on information granules and fuzzy clustering,” Engineering Applications of Artificial Intelligence, 41 (2015) 17–24.

[3] Hahsler, Michael, “Introduction to arules – A computational environment for mining association rules and frequent item sets, ” 2005, Journal of Statistical Software.

[4] Han, J., & Kamber, M, “Data mining: Concepts and techniques (2nd ed.),” San Francisco, CA: Morgan Kaufmann, 2006.

[5] Wang, J. C., Rygielski, C., & Yen, D. C, “Data mining techniques for customer relationship management,” Technology in Society, 24(4), 483–502, 2002.

[6] Herawan, T., Vitasari, P., & Abdullah, Z. “Mining interesting association rules of students suffering study anxieties using SLP-growth algorithm,” International Journal of Knowledge and Systems Science, 3(2), 24–41, 2012.

[7] M. Delgado, N. Marin, M.J. Martin-Bautista, D. Sanchez, et al, “Soft computing for information processing and analysis,” Springer Berlin Heidelberg , vol. 164 of the series studies in fuzziness and soft Computing, 2005, pp 351-373.

[8] J. Ruan, X. Wang, Y. Shi, “Developing fast predictors for large-scale time series using fuzzy granular support vector machines,” Applied Soft Computing 13 (2013) 3981–4000.

[9] L. A. Zadeh. Fuzzy sets and information granularity. Advances in Fuzzy Set Theory and Application, 1, (1979) 3–18.

[10] Chengzhong Luo, “Introduction to fuzzy sets,” Beijing Normal University Press, vol. 1, pp. 121–124, August 2005.

[11] M.A. Erceg, “Metric spaces in fuzzy set theory,” Journal of Mathematical Analysis and its Applications, 69, (1979) 205–230.

[12] B. Bede, “Mathematics of fuzzy sets and fuzzy logic,” Springer Heidelberg New York Dordrecht London, (2013).

[13] Xiyang Yang, Fusheng Yu, Witold Pedrycz, “Long-term forecasting of time series based on linear fuzzy information granules and fuzzy inference system,” International Journal of Approximate Reasoning, February 2017, Vol.81, pp.1-27.

[14] X. Yan, X. Su. “Linear regression analysis: theory and computing,” Singapore: World Scientific Publishing CO. Pte. Ltd; (2009) 1–2.

[15] Nayak J, Naik B, Behera HS, “Fuzzy c-means (FCM) clustering algorithm: a decade review from 2000 to 2014,” Comput Intell Data Min 2:133–149 Springer India, (2015).

[16] Jianjiang Lu, “Research and Application of Fuzzy Association Rules,” Science Press, January 2008, pp 30-33.

[17] This time series is from: https://datamarket.com/data/set/235d/mean-daily-temperature-fisher-river-near-dallas-jan-01-1988-to-dec-31-1991