*International Journal of Innovative Computing, Information and Control*
Volume **8**, Number **3(B)**, March **2012**

ICIC International ©2012 ISSN 1349-4198
pp. 2113–2124

# CLUSTERING PERIODIC FREQUENT PATTERNS USING FUZZY STATISTICAL PARAMETERS

Fokrul Alom Mazarbhuiya[1] and Muhammad Abulaish[2]

[1]College of Computer Science
King Khalid University
P.O. Box 3236, Abha, Saudi Arabia
fokrul_2005@yahoo.com

[2]Center of Excellence in Information Assurance
King Saud University
P.O. Box 92144, Riyadh 11653, Saudi Arabia
mAbulaish@ksu.edu.sa

Received December 2010; revised April 2011

Abstract. *Frequent pattern mining from super-market transaction datasets is a well-stated data mining problem and consequently there are number of approaches including association rule mining to deal with this problem. However, super-market transaction datasets are generally temporal in the sense that, when a transaction happens in a super-market, the time of the transaction is also recorded. A number of techniques have been proposed to find frequent itemsets from temporal datasets in which each itemset is associated with a list of time intervals in which it is frequent. Considering the time of transactions as calendar dates, there may exist various types of periodic patterns viz. yearly, quarterly, monthly, daily, etc. And, if the time intervals associated with a periodic itemset are kept in a compact manner then it turns out to be a fuzzy time interval for which the set superimposition method can be used. In this paper, we propose an agglomerative hierarchical clustering algorithm to find clusters among the periodic itemsets. Since the fuzzy number is invariant with respect to shifting, we define similarity measure using the variance of fuzzy intervals associated with frequent itemsets. The efficacy of the proposed method is established through experimentation on real datasets.*

**Keywords:** Data mining, Clustering, Temporal patterns, Locally frequent itemset, Set superimposition, Fuzzy time-interval

1. **Introduction.** Clustering is one of the well-known data mining problems which follows unsupervised learning approach and it is very useful for the discovery of data distribution and patterns in the datasets with unknown class-labels. The goal of the clustering process is to discover both the dense and sparse regions in a dataset. There are two main approaches to clustering: *hierarchical clustering* and *partitioning clustering*. In hierarchical clustering, the dataset is divided into a sequence of partitions, in which each partition is nested into the next partition in the sequence. The hierarchical clustering creates a *hierarchy* of clusters from small to big or big to small and consequently it is termed as *agglomerative* or *divisive* clustering techniques respectively. Clustering of numerical data has been studied in the past [8]. However, in real life, we come across datasets that contain different types of data such as binary, categorical, spatial, ordinal, temporal or mixture of these. During the last few years, a number of new and interesting algorithms for clustering categorical and spatial data have also been proposed [5, 15, 16, 18, 20].

Association rule mining is another data mining problem which focuses on deriving associations among data. The association rule mining problem was formulated by Agrawal et al. [14] and is often referred to as *market-basket analysis* problem. Since their introduction, the *frequent itemset mining* and *association rule mining* problems have received a great deal of attention. During the last few decades, hundreds of research papers have been published presenting new algorithms or improvements of existing algorithms to solve the association rule mining problem more efficiently [21]. Mining association rules from *temporal dataset* is also an interesting data-mining problem and recently it has received a great deal of attention. In [9], Ale et al. have proposed a method of extracting association rules which hold through out the life-span of an itemset where the life-span of an itemset is defined as the time-period between the first transaction and last transaction containing the itemset. Here life-span of an itemset may not be the same as that of the dataset. However, there may be items that appear in the transactions for a short time period, but within that period, they appear frequently and then they disappear for a long period and appear again. For such items, when the support values are calculated with respect to their life-span, they may become infrequent. Another problem is that even if an itemset is frequent within its life-span, the density of occurrences of the items in the transactions may vary. Within the lifespan, there may be time-slots when the density of occurrences is high and some time-slots where the density is very low. Identifying such dense time-slots will certainly provide knowledge about the dataset and the items concerned. In [1], the work proposed by Ale and Rossi [9] is extended by incorporating time-gap between two consecutive transactions containing an item to solve some of these issues.

The algorithm proposed in [1] outputs all locally frequent itemsets along with the list of time-intervals. Each frequent itemset is associated with a list of time-intervals where it is frequent. The list of time-intervals associated with frequent itemsets may satisfy some interesting properties. For example, considering the time-stamps as calendar dates, we can define periodic patterns with different levels of specificities. All such patterns are discussed in detail by the same authors in [2]. While extracting periodicity of a frequent pattern, if the associated time-intervals have large overlapping, then the intervals can be stored using *set superimposition* [7], which turns out to be fuzzy interval. For example, in some countries, the winter season (when the temperature goes down) starts around November/December and continues till February/March. These cut-off times are not always the same. It varies from year to year, but even then these durations never have empty intersections. Under this assumption and the assumption that the items are uniformly distributed in the period in which they are frequent (which is reasonable as we are considering locally frequent sets, ignoring the time periods in which an item set appears rarely), our representation scheme using set superimposition gives fuzzy time intervals. In this way, we can have some periodic patterns associated with fuzzy time intervals describing their periods.

In this paper, we propose an agglomerative hierarchical clustering algorithm to find clusters among such periodic patterns. As the variance of a fuzzy number is invariant with respect to translation, it can be used to define the similarity measures between clusters consisting of periodic patterns. The objective of the paper is threefold. First, it defines similarity between pair of periodic patterns having fuzzy time-intervals describing their periods as ratio of difference of variances of fuzzy time-intervals to the sum of the same. If the similarity value is less than a pre-assigned threshold, then the corresponding patterns will be similar and will belong to the same clusters, otherwise they will belong to different clusters. If the similarity value is either 0 or 1, then the corresponding patterns are exactly similar or exactly dissimilar respectively. Secondly, it defines the similarity between pair of clusters consisting of similar periodic patterns as the ratio of difference of the average of variances of fuzzy time-intervals of the similar patterns belonging to each cluster to the sum of the same. Then, a *merge* function is defined in terms of the similarity. If the value of the similarity function is less than a pre-assigned threshold, then the corresponding cluster pairs are similar and they will be merged using *merge* function to form a larger cluster. Finally, in this paper, we present an algorithm for the clustering of periodic patterns. The algorithm is agglomerative hierarchical and is in-line with the one proposed in [10].

The rest of the paper is organized as follows. Section 2 presents a brief review of the existing clustering algorithms. In Section 3, we present some basic definitions and results related to periodic patterns and fuzzy time-intervals. The proposed agglomerative clustering algorithm is presented in Section 4. In Section 5, we discuss our experimental setup and results. Finally, we conclude the paper with possible future enhancements of the proposed work in Section 6.

2. **Related Work.** In this section, we present a brief review of the existing research findings related to our work. In [11], an algorithm for clustering categorical data has been proposed. Using a similarity function to measure the similarity between pairs of points and a user defined threshold $\theta$, pairs of points for which the value of the similarity function is greater than or equal to $\theta$ are considered as neighbors. Starting with each point in its own cluster, the algorithm repeatedly merges clusters to form larger clusters. If two distinct data points are neighbors, then the corresponding two clusters are merged. The process ends when no pair of distinct neighbors belongs to two distinct clusters. The algorithm BIRCH proposed in [17] is also a hierarchical agglomerative clustering algorithm. It is designed for clustering large datasets using limited amount of memory and it assumes the dataset to be numeric in nature. Like BIRCH, the ROCK algorithm introduced by Guha et al. [16] is also an agglomerative hierarchical clustering, but it can be applied on categorical data. During the last few years the concept of fuzzy sets [19] has been widely used in different areas including cluster analysis and pattern recognition [22, 23, 24]. In [10], the author has proposed an algorithm for clustering categorical data using a fuzzy set based approach. The algorithm is agglomerative in which the clusters are represented as fuzzy sets and the similarity measure is defined in terms of corresponding fuzzy set representations. The concept of finding associations among data has also attracted a large number of researchers. An efficient algorithm for the discovery of association rule is presented in [13], which was then followed by refinements, generalizations, extensions and improvements. The association rules generation process is also extended to incorporate temporal aspects. In [9], an algorithm for discovery of temporal association rules is described where for each item (which is extended to itemset) a lifetime or life-span is defined as the time gap between the first and the last occurrences of the transactions containing the item. Supports of items are calculated only during its life-span. Thus each association rule is associated with a time frame corresponding to the lifetime of the items participating in the rule. In [1], the works proposed in [9] is extended by incorporating time-gap between two consecutive transactions containing an item. The algorithm [1] gives all locally frequent itemsets along with the lists of time intervals. Each frequent itemset is associated with a list of time intervals where it is frequent. In order to compute the periodicity of such frequent itemsets if the associated time intervals have large overlap, then a method of redefining the intervals is proposed in [7], which turns out to be fuzzy intervals. In this paper, our focus is to develop an agglomerative hierarchical clustering method to cluster frequent temporal patterns using fuzzy time intervals associated with them.

3. **Definitions and Results.** In this section, we present a summarized view of some basic concepts, definitions and results on which our proposed algorithm is based.

3.1. **Fuzziness.** Let $E$ be the universe of discourse. A fuzzy set $A$ in $E$ is characterized by a membership function $A(x)$ lying in $[0, 1]$. For any $x \in E$, $A(x)$ represents the grade of membership of $x$ in $A$. Thus, a fuzzy set $A$ is defined as given in Equation (1).

$$A = \{(x, A(x)), x \in E\} \tag{1}$$

A fuzzy set $A$ is said to be normal if $A(x) = 1$ for at least one $x \in E$. An $\alpha$-cut [6] of a fuzzy set $A$ is represented by $A_\alpha$ and defined as an ordinary set of elements with membership grade greater than or equal to a threshold $\alpha$, $0 \leq \alpha \leq 1$. Thus an $\alpha$-cut $A_\alpha$ of a fuzzy set $A$ is characterized by: $A_\alpha = \{x \in E; A(x) \geq \alpha\}$. A fuzzy set is said to be *convex* if all its $\alpha$-cuts are convex sets. A fuzzy number is a convex normalized fuzzy set $A$ defined on the real line $R$ such that $(i)$ there exists an $x_0 \in R$ such that $A(x_0) = 1$, and $(ii)$ $A(x)$ is piecewise continuous. Thus a fuzzy number can be thought of as containing the real numbers within some interval to varying degrees. Fuzzy intervals are special fuzzy numbers satisfying the following two conditions: $(i)$ there exists an interval $[a, b] \subset R$ such that $A(x_0) = 1$ for all $x_0 \in [a, b]$, and $(ii)$ $A(x)$ is piecewise continuous. A fuzzy interval can be thought of as a fuzzy number with a flat region. A fuzzy interval $A$ is denoted by $A = [a, b, c, d]$ with $a < b < c < d$, where $A(a) = A(d) = 0$ and $A(x) = 1$ for all $x \in [b, c]$. $A(x)$ for all $x \in [a, b]$ is known as *left reference function* and $A(x)$ for all $x \in [c, d]$ is known as the *right reference function.* The left reference function is non-decreasing and the right reference function is non-increasing [4].

3.2. **Set superimposition.** When we overwrite, the overwritten portion looks darker for obvious reasons. The set operation union does not explain this phenomenon, i.e.,

$$A \cup B = (A - B) \cup (A \cap B) \cup (B - A) \tag{2}$$

and, in $(A \cap B)$ the elements are represented once only. In [7], Baruah introduced an operation called set superimposition which is denoted by $(S)$. According to his definition, if a set $A$ is superimposed over a set $B$ or $B$ is superimposed over $A$, we have

$$A(S)B = (A - B)(+)(A \cap B)^{(2)}(+)(B - A) \tag{3}$$

where $(A \cap B)^{(2)}$ are the elements of $(A \cap B)$ represented twice, and $(+)$ represents union of disjoint sets.

For illustration purpose, let $A = [a_1, b_1]$ and $B = [a_2, b_2]$ are two real intervals such that $A \cap B \neq \phi$, and we have to find a superimposed portion. It can be seen from Equation (3) that

$$[a_1, b_1]\,(S)\,[a_2, b_2] = \left[a_{(1)}, a_{(2)}\right](+)\left[a_{(2)}, b_{(1)}\right]^{(2)}(+)\left[b_{(1)}, b_{(2)}\right] \tag{4}$$

where $a_{(1)} = \min(a_1, a_2)$, $a_{(2)} = \max(a_1, a_2)$, $b_{(1)} = \min(b_1, b_2)$ and $b_{(2)} = \max(b_1, b_2)$. Equation (4) explains why if two line segments are superimposed, the common portion looks doubly dark [6]. The identity (4) is called fundamental identity of superimposition of intervals. Now, let $[a_1, b_1]^{(1/2)}$ and $[a_2, b_2]^{(1/2)}$ be two fuzzy sets with constant membership value 1/2 everywhere (i.e., equi-fuzzy intervals with membership value 1/2). Applying identity (4) on the two equi-fuzzy intervals, we can write

$$[a_1, b_1]^{(1/2)}\,(S)\,[a_2, b_2]^{(1/2)} = \left[a_{(1)}, a_{(2)}\right]^{(1/2)}(+)\left[a_{(2)}, b_{(1)}\right]^{(1)}(+)\left[b_{(1)}, b_{(2)}\right]^{(1/2)} \tag{5}$$

To explain this let us consider the fuzzy intervals $[1, 5]^{(1/2)}$ and $[3, 7]^{(1/2)}$ with constant membership value 1/2 as shown in Figures 1$(a)$ and 1$(b)$. Here $[1, 5] \cap [3, 7] = [3, 5] \neq \phi$ If we apply superimposition on the intervals then the superimposed interval will consist of $[1, 3]^{(1/2)}$, $[3, 5]^{(1)}$ and $[5, 7]^{(1/2)}$ as shown in Figure 1$(c)$. Here, the membership of the

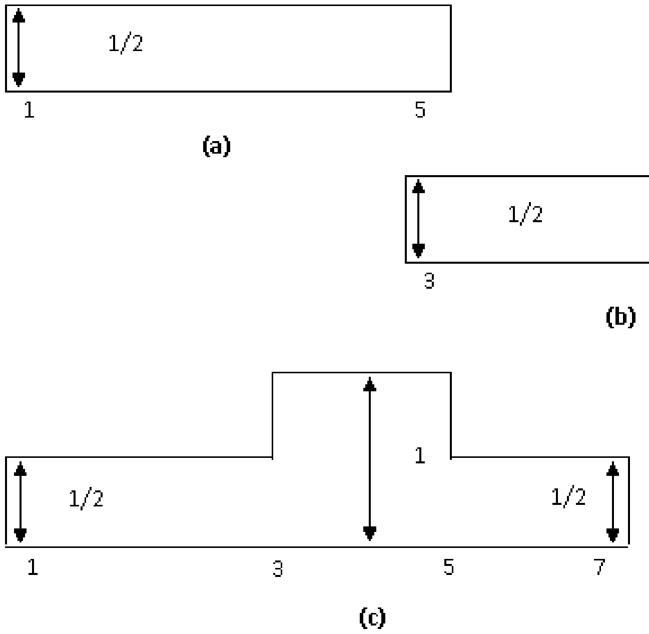


Figure 1: Set superimposition example. (a) Fuzzy interval $[1,5]^{(1/2)}$ with constant membership value 1/2. (b) Fuzzy interval $[3,7]^{(1/2)}$ with constant membership value 1/2. (c) Superimposed interval consisting of fuzzy intervals $[1,3]^{(1/2)}$, $[3,5]^{(1)}$ and $[5,7]^{(1)}$ with constant membership values 1/2, 1 and 1/2 respectively.

interval $[3,5]$ is 1 due to its double representation. Now, let $[x_i, y_i]$, $i = 1, 2, \cdots, n$ are $n$ real intervals such that $\bigcap_{i=1}^{n}[x_i, y_i] \neq \phi$. Generalizing Equation (5), we get

$$\begin{aligned}&[x_1,y_1]^{(1/n)}(S)[x_2,y_2]^{(1/n)}(S)\cdots(S)[x_n,y_n]^{(1/n)}\\ =&[x_{(1)},x_{(2)}]^{(1/n)}(+)[x_{(2)},x_{(3)}]^{(2/n)}(+)\cdots(+)[x_{(r)},x_{(r+1)}]^{(r/n)}\\ &(+)\cdots(+)[x_{(n)},y_{(1)}]^{(1)}(+)[y_{(1)},y_{(2)}]^{((n-1)/n)}(+)\cdots(+)[y_{(n-r)},y_{(n-r+1)}]^{(r/n)}\\ &(+)\cdots(+)[y_{(n-2)},y_{(n-1)}]^{(2/n)}(+)[y_{(n-1)},y_{(n)}]^{(1/n)}\end{aligned} \tag{6}$$

In Equation (6), the sequence $\{x_{(i)}\}$ is formed of the sequence $\{x_i\}$, $i = 1, 2, \cdots, n$ in ascending order of magnitude and similarly $\{y_{(i)}\}$ is formed of the sequence $\{y_i\}$, $i = 1, 2, \cdots, n$ in ascending order of magnitude.

**Lemma 3.1.** *(The Glivenko-Cantelli Lemma of Order Statistics). Let $X = (X_1, X_2, \cdots, X_n)$ and $Y = (Y_1, Y_2, \cdots, Y_n)$ be two random vectors, and $(x_1, x_2, \cdots, x_n)$ and $(y_1, y_2, \cdots, y_n)$ be two particular realizations of $X$ and $Y$ respectively. Assume that the sub-$\sigma$ fields induced by $X_k$, $k = 1, 2, \cdots, n$ are identical and independent. Similarly, assume that the sub-$\sigma$ fields induced by $Y_k$, $k = 1, 2, \cdots, n$ are also identical and independent. Let $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ be the values of $x_1, x_2, \cdots, x_n$, and $y_{(1)}, y_{(2)}, \cdots, y_{(n)}$ be the values of $y_1, y_2, \cdots, y_n$ arranged in ascending order. For $X$ and $Y$ if the empirical probability distribution functions $\phi_1(x)$ and $\phi_2(y)$ are defined as in Equations (7) and (8) respectively, then the Glivenko-Cantelli Lemma of order statistics states that the mathematical expectation of the empirical probability distributions would be given by the respective theoretical* *probability distributions* [12].

$$\phi_1(x) = \begin{cases} 0 & \textit{if } x < x_{(1)} \\ \dfrac{(r-1)}{n} & \textit{if } x_{(r-1)} \le x \le x_{(r)} \\ 1 & \textit{if } x \ge x_{(n)} \end{cases} \tag{7}$$

$$\phi_2(y) = \begin{cases} 0 & \textit{if } y < y_{(1)} \\ \dfrac{(r-1)}{n} & \textit{if } y_{(r-1)} \le y \le y_{(r)} \\ 1 & \textit{if } y \ge y_{(n)} \end{cases} \tag{8}$$

Now, let $X_k$ is random in the interval $[a, b]$ and $Y_k$ is random in the interval $[b, c]$ so that $P_1(a, x)$ and $P_2(b, y)$ are the probability distribution functions followed by $X_k$ and $Y_k$ respectively. Then in this case Glivenko-Cantelli Lemma gives:

$$E\left[\phi_1(x)\right] = P_1(a, x),\ a \le x \le b \text{ and } E\left[\phi_2(y)\right] = P_1(b, y),\ b \le y \le c \tag{9}$$

It can be observed that in Equation (6) the membership values of $[x_{(r)}, x_{(r+1)}]^{(r/n)}$, $r = 1, 2, \cdots, n-1$ look like empirical probability distribution function $\phi_1(x)$ and the membership values of $[y_{(n-r)}, y_{(n-r+1)}]^{(r/n)}$, $r = 1, 2, \cdots, n-1$ look like the values of empirical complementary probability distribution function or empirical survival function $[1 - \phi_2(y)]$. Therefore, if $A(x)$ is the membership function of an L-R fuzzy number $A = [a, b, c]$. We get from Equation (9)

$$A(x) = \begin{cases} P_1(a, x) & \text{if } a \le x \le b \\ 1 - P_2(b, x) & \text{if } b \le x \le c \end{cases} \tag{10}$$

Thus it can be seen that $P_1(x)$ can indeed be the Dubois-Prade left reference function and $(1 - P_2(x))$ can be the Dubois-Prade right reference function [4]. Baruah [7] has shown that if a possibility distribution is viewed in this way, two probability laws can, indeed, give rise to a possibility law.

**Definition 3.1.** *(Possibilistic mean and possibilistic variance of a fuzzy number). Let $F$ be a family of fuzzy numbers and $A$ be a fuzzy number belonging to $F$. Let $A_\alpha = [a_1(\alpha), a_2(\alpha)]$, $\alpha \in [0, 1]$ be an $\alpha$-cut of $A$. Carlsson and Fuller* [3] *defined possibilistic variance of fuzzy number $A \in F$ as:*

$$Var(A) = \frac{1}{2}\int_0^1 \alpha(\alpha_2(\alpha) - \alpha_1(\alpha))^2 d\alpha \tag{11}$$

In the following paragraph, we present a proof of the theorem given by Carlsson and Fuller [3].

**Theorem 3.1.** *The variance of a fuzzy number is invariant to shifting.*

**Proof:** Let $A \in F$ be a fuzzy number and let $\theta$ be a real number. If $A$ is shifted by value $\theta$, then we get a fuzzy number, denoted by $B$, satisfying the property $B(x) = A(x - \theta)$ for all $x \in R$. From the relationship $B_\alpha = [a_1(\alpha) + \theta, a_2(\alpha) + \theta]$, we may write

$$\begin{aligned} Var(B) &= \frac{1}{2}\int_0^1 \alpha((\alpha_2(\alpha) + \theta) - (\alpha_1(\alpha) + \theta))^2 d\alpha \\ &= \frac{1}{2}\int_0^1 \alpha(\alpha_2(\alpha) - \alpha_1(\alpha))^2 d\alpha \\ &= Var(A) \end{aligned}$$

**Definition 3.2.** *(Similarity measure between pairs of patterns having fuzzy time intervals as their periods). Let $A_1$ and $A_2$ be two patterns with fuzzy time intervals $T_1$ and $T_2$ respectively. The similarity measure between $A_1$ and $A_2$ is represented as $sim(A_1, A_2)$ and defined using Equation (12) in which $var(T_1)$ is the variance of the fuzzy time interval $T_1$ associated with $A_1$, $var(T_2)$ is the variance of the fuzzy time interval $T_2$ associated with $A_2$, and $|\ |$ is the absolute value function.*

$$sim(A_1, A_2) = \left| \frac{var(T_1) - var(T_2)}{var(T_1) + var(T_2)} \right| \tag{12}$$

We consider two patterns as similar if and only if value of their *sim* function is less than or equal to a pre-assigned threshold value, otherwise they will be dissimilar. For a value 0 they will be precisely similar and that for 1 they will be precisely dissimilar.

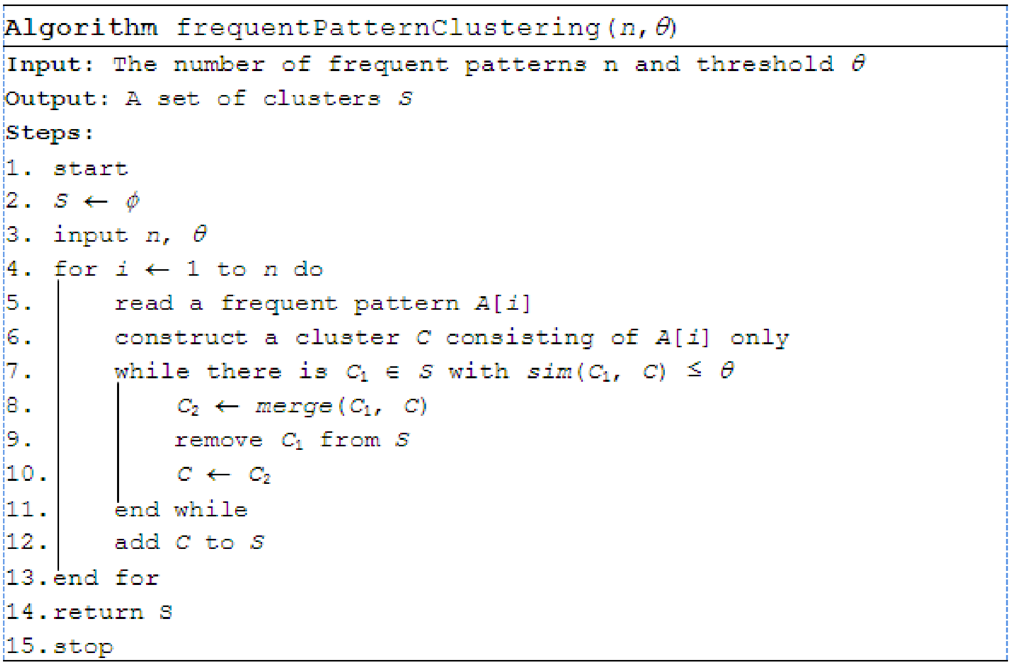
**Definition 3.3.** *(Similarity of pairs of clusters containing similar patterns). Let $C_1$ and $C_2$ be two clusters and let $C_1$ consist of similar patterns say $\{A[i];\ i = 1, 2, \cdots, n_1\}$ and $C_2$ consists of similar patterns say $\{B[i];\ i = 1, 2, \cdots, n_2\}$. The similarity between $C_1$ and $C_2$ is defined using Equations (13) in which $D_1 = \sum_{i=1}^{n_1} var(T[i])/n_1$ is the average of the variances of $\{T[i];\ i = 1, 2, \cdots, n_1\}$ that are associated with the similar patterns $\{A[i];\ i = 1, 2, \cdots, n_1\}$ of $C_1$ and $D_2 = \sum_{i=1}^{n_2} var(T'[i])/n_2$ is the average of the variances of $\{T'[i];\ i = 1, 2, \cdots, n_2\}$ that are associated with the similar patterns $\{B[i];\ i = 1, 2, \cdots, n_2\}$ of $C_2$.*

$$sim(C_1, C_2) = \left| \frac{D_1 - D_2}{D_1 + D_2} \right| \tag{13}$$

**Definition 3.4.** *(Merger of clusters). Let $C_1$ and $C_2$ be two clusters having $n_1$ and $n_2$ patterns respectively. Let $C$ be the cluster obtained by merging $C_1$ and $C_2$. Then the merge function is defined as merge $(C_1, C_2) = C_1 \cup C_2$, if and only if $sim(C_1, C_2) \leq \theta$, where $\theta$ is a pre-defined threshold value.*

## 4. Proposed Algorithm.

In this section, we present the proposed clustering algorithm based on the concepts discussed in the previous section. For the proposed algorithm, all periodic patterns having fuzzy time intervals describing their periods serves as input data. The fuzzy time intervals are obtained by the methods discussed in Section 3. Since the variance of fuzzy intervals are invariant to shifting, two periodic patterns having the same value of variance for the fuzzy intervals describing their periods can be considered to be similar. Considering frequent patterns along with fuzzy time intervals describing their periods (each pattern is associated with exactly one fuzzy time interval), we want to find clusters among frequent patterns such that all similar frequent patterns are grouped in the same cluster. The similarity between two patterns is defined in terms of variance of the fuzzy time intervals associated with them, i.e., two patterns having fuzzy time intervals $T_1$ and $T_2$ are similar if and only if the value of the corresponding *sim* function (defined in Equation (12)) is less than a pre-defined threshold. In order to start the clustering process, each pattern is assigned to a separate cluster. Thereafter, for each pair of clusters the similarity value is calculated and merge function is applied (to generate a new bigger cluster), if the similarity value is within the threshold. The process of merging continues till no merger of clusters is possible or there is only one cluster at the top. In this way, the process to generate clusters is hierarchical-agglomerative. Algorithm frequent Pattern Clustering, shown in Figure 2, presents the pseudo code for the proposed algorithm.



```
Algorithm frequentPatternClustering($n$, $\theta$)
Input: The number of frequent patterns n and threshold $\theta$
Output: A set of clusters $S$
Steps:
start
$S \leftarrow \phi$
input $n$, $\theta$
for $i \leftarrow 1$ to $n$ do
    read a frequent pattern $A[i]$
    construct a cluster $C$ consisting of $A[i]$ only
    while there is $C_1 \in S$ with $sim(C_1, C) \leq \theta$
        $C_2 \leftarrow merge(C_1, C)$
        remove $C_1$ from $S$
        $C \leftarrow C_2$
    end while
    add $C$ to $S$
end for
return S
stop
```

Figure 2: Algorithm to cluster frequent patterns with fuzzy time intervals

5. **Experimental Setting and Results.** For experimental purpose, we have used a synthetic dataset `T10I4D100K`, available from `FIMI`[1] website. A summarized view of the dataset describing the number of items, the number of transactions, and the minimum, maximum and average length of transactions is presented in Table 1. Since the dataset is non-temporal it cannot be used in its current form for our experimentation. Therefore, a program was written to incorporate temporal features in the dataset. The program takes as input a starting date and two values for the minimum and the maximum number of transactions per day. A number between these two limits are selected at random and that many consecutive transactions are labeled with the same date to reflect the fact that many transactions have taken place on that day. This process starts from the first transaction and continue to the end by marking the transactions with consecutive dates (assuming that the market remains open on all week days). The process is repeated for first 10000, 20000, 30000, 40000, 50000, 60000 transactions and then for whole dataset containing 100000 transactions to generate the datasets of different sizes. We have given the maximum number of transactions and minimum number of transactions in such a way that the lifetime of the datasets for each size is almost one year. A detail description of the temporal datasets obtained through this process is presented in Table 2.

Table 1: T10I4D100K dataset characteristics

| Dataset | # Items | # Transactions | $min\|T\|$ | $max\|T\|$ | $avg\|T\|$ |
|---|---|---|---|---|---|
| T10I4D100K | 942 | 100000 | 4 | 77 | 39 |

[1] http://fimi.cs.helsinki.fi/data/

Table 2: Temporal datasets and their characteristics

| Dataset | # Transactions | $Min\|T\|$ | $Max\|T\|$ | Start Date | End Date |
|---|---|---|---|---|---|
| T1 | 10000 | 25 | 30 | 1-1-2000 | 29-12-2000 |
| T2 | 20000 | 50 | 60 | 1-1-2000 | 26-12-2000 |
| T3 | 30000 | 75 | 90 | 1-1-2000 | 02-01-2001 |
| T4 | 40000 | 100 | 120 | 1-1-2000 | 26-12-2000 |
| T5 | 50000 | 125 | 150 | 1-1-2000 | 31-12-2000 |
| T6 | 60000 | 150 | 180 | 1-1-2000 | 03-01-2001 |
| T7 | 100000 | 250 | 300 | 1-1-2000 | 20-12-2000 |

Table 3: A partial list of periodic itemsets and their time-intervals

| Periodic Itemset | Time Intervals |
|---|---|
| 25 | [2-1-2000, 10-1-2000], [4-3-2000, 15-3-2000], [5-4-2000, 12-4-2000] |
| 31 | [3-1-2000, 13-1-2000], [1-3-2000, 13-3-2000], [3-4-2000, 12-4-2000], [5-7-2000, 14-7-2000] |
| 25, 31 | [4-1-2000, 10-1-2000], [5-3-2000, 12-3-2000], [6-4-2000, 10-4-2000] |
| 50 | [6-4-2000, 20-4-2000], [6-6-2000, 18-6-2000], [6-7-2000, 16-7-2000], [6-10-2000, 15-10-2000] |
| 89 | [12- 3-2000, 23-3-2000] |
| 112 | [8-5-2000, 15-5-2000], [15-7-2000, 23-7-2000], [12-10-2000, 20-10-2000] |

Table 4: A partial list of periodic itemsets and their fuzzy time-intervals

| Periodic Itemset | Superimposed Intervals | Fuzzy Intervals |
|---|---|---|
| 25 | $[2,4]^{(1/3)}[4,5]^{(2/3)}[5,10]^{(1)}[10,12]^{(2/3)}[12,15]^{(1/3)}$ | [2, 5, 10,15] |
| 31 | $[1,3]^{(1/4)}[3,3]^{(2/4)}[3,5]^{(3/4)}[5,12]^{(1)}[12,13]^{(3/4)}$ $[13,13]^{(2/4)}[13,14]^{(1/4)}$ | [1, 5, 12, 14] |
| 25, 31 | $[4,5]^{(1/3)}[5,6]^{(2/3)}[6,10]^{(1)}[10,10]^{(2/3)}[10,12]^{(1/3)}$ | [4, 6, 10, 12] |
| 50 | $[6,6]^{(1/4)}[6,6]^{(2/4)}[6,6]^{(3/4)}[6,15]^{(1)}[15,16]^{(3/4)}$ $[16,18]^{(2/4)}[18,20]^{(1/4)}$ | [6, 6, 15, 20] |
| 89 | $[2,3]^{(1/3)}[3,5]^{(2/3)}[5,10]^{(1)}[10,15]^{(2/3)}[15,20]^{(1/3)}$ | [2, 5, 10, 20] |
| 112 | $[8,12]^{(1/3)}[12,15]^{(2/3)}[15,15]^{(1)}[15,20]^{(2/3)}[20,23]^{(1/3)}$ | [8, 15, 15, 23] |

In order to get periodic-patterns where period of each itemset is described by fuzzy time interval, we have applied our proposed calendar-based approach presented in [2]. Thereafter, we apply the concept of superimposition on non-empty intervals to find fuzzy interval for each periodic itemset. In [2], there were some restrictions in forming fuzzy intervals, i.e., the intervals to be superimposed must have large overlaps and the number of intervals must be greater than or equal to some threshold. Both of these restrictions are relaxed in this paper, i.e., we consider all overlapping intervals associated with a periodic itemset to form fuzzy intervals and similarly, we consider all frequent itemsets that are frequent in any one interval. A partial list of periodic itemsets along with their time periods where they are frequent is given in Table 3. In order to find the monthly periodic

Table 5: Clustering results along with the number of misclassified itemsets for different set of transactions

| Dataset | Max. No. of Itemsets | # Clusters Obtained | # Itemsets Misclassified |
|---|---|---|---|
| T1 | 123 | 10 | 3 |
| T2 | 200 | 12 | 3 |
| T3 | 235 | 13 | 2 |
| T4 | 312 | 15 | 2 |
| T5 | 350 | 20 | 1 |
| T6 | 400 | 24 | 1 |
| T7 | 942 | 26 | 0 |

itemsets, we remove the year and month from time hierarchy. Thereafter, we apply the superimposition of intervals to get fuzzy intervals for the frequent itemsets. Table 4 shows the superimposed and fuzzy intervals for the frequent itemsets of Table 3.

Thereafter, we have applied the proposed agglomerative-hierarchical algorithm to find clusters among the periodic patterns. For threshold value ($\theta = 0.4$), the clustering results along with the number of misclassified itemsets obtained from the dataset mentioned in Table 2 is presented in Table 5. It can be observed from Table 5 that with increasing number of transactions in the datasets the number of misclassified itemsets are less.

6. **Conclusion and Future Work.** In this paper, we have presented an agglomerative-hierarchical clustering algorithm to find clusters among periodic patterns with fuzzy time intervals. The algorithm starts with as many clusters as the periodic patterns having fuzzy time intervals. Then, the pairs of clusters are merged if their similarity value is less than a pre-defined threshold. The process continues till a specified number of clusters is obtained or there is no two patterns having similarity value less than the threshold and belongs to two different clusters. We have also presented a similarity measure defined in terms of variances of the fuzzy time intervals associated with the corresponding periodic patterns where fuzzy time intervals are obtained using a method based on set superimposition.

Although, we have used the agglomerative-hierarchical algorithm for clustering purpose, any other clustering algorithm can be applied provided the similarity measure is properly defined. Moreover, instead of variance other statistical parameters can be used to define similarity measure in future.

**Acknowledgment.** The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

## REFERENCES

[1] A. K. Mahanta, F. A. Mazarbhuiya and H. K. Baruah, Finding locally and periodically frequent sets and periodic association rules, *LNCS*, vol.3776, pp.576-582, 2005.

[2] A. K. Mahanta, F. A. Mazarbhuiya and H. K. Baruah, Finding calendar-based periodic patterns, *Pattern Recognition Letters*, vol.29, no.9, pp.1274-1284, 2008.

[3] C. Carlsson and R. Fuller, On possibilistic mean value and variance of fuzzy numbers, *Fuzzy Sets and Systems*, vol.122, pp.315-326, 2001.

[4] D. Dubois and H. Prade, Ranking fuzzy numbers in the setting of possibility theory, *Information Science*, vol.30, pp.183-224, 1983.

[5] D. Gibson, J. Kleinberg and P. Raghavan, Clustering categorical data: An approach based on dynamical systems, *Proc. of the 24th Int'l. Conf. on Very Large Databases*, New York, USA, pp.311-323, 1998.

[6] G. Q. Chen, C. Lee Samuel and S. H. Eden Yu, Application of fuzzy set theory to economics, *Advances in Fuzzy Sets, Possibility Theory, and Applications*, pp.277-305, 1983.

[7] H. K. Baruah, Set superimposition and its application to the theory of fuzzy sets, *Journal of Assam Science Society*, vol.10, no.1-2, pp.25-31, 1999.

[8] J. A. Hartigan, *Clustering Algorithms*, John Wiley & Sons, New York, NY, USA, 1975.

[9] J. M. Ale and G. H. Rossi, An approach to discovering temporal association rules, *Proc. of ACM Symposium on Applied Computing*, 2000.

[10] M. Dutta and A. K. Mahanta, An algorithm for clustering large categorical databases using a fuzzy set based approach, *Proc. of the 17th Australian Joint Conf. on Artificial Intelligence*, Cairns, Australia, 2004.

[11] M. Dutta, A. K. Mahanta and M. Mazumder, An algorithm for clustering of categorical data using concept of neighours, *Proc. of the 1st National Workshop on Soft Data Mining and Intelligent Systems*, Tezpur University, India, pp.103-105, 2001.

[12] M. Loeve, *Probability Theory*, Springer Verlag, New York, NY, USA, 1977.

[13] R. Agrawal and R. Srikant, Fast algorithms for mining association rules, *Proc. of the 20th Int'l. Conf. on Very Large Databases*, Santiago, Chile, 1994.

[14] R. Agrawal, T. Imielinski and A. N. Swami, Mining association rules between sets of items in large databases, *ACM SIGMOD Record*, vol.22, no.2, pp.207-216, 1993.

[15] R. T. Ng and J. Han, Efficient and effective clustering methods for spatial data mining, *Proc. of the 20th Int'l. Conf. on Very Large Databases*, Santiago, Chile, pp.144-155, 1994.

[16] S. Guha, R. Rastogi and K. Shim, ROCK: A robust clustering algorithm for categorical attributes, *Proc. of the IEEE Int'l. Conf. on Data Engineering*, Sydney, Australia, pp.512-521, 1999.

[17] T. Zhang, R. Ramakrishnan and M. Livny, BIRCH: An efficient data clustering method for very large databases, *Proc. of the ACM SIGMOD Conf. on Management of Data*, Canada, pp.103-114, 1996.

[18] V. Ganti, J. Gehrke and R. Ramakrishnan, CACTUS-clustering categorical data using summaries, *Proc. of the Int'l. Conf. on Knowledge Discovery and Data Mining*, San Diego, CA, USA, pp.73-83, 1999.

[19] L. A. Zadeh, Fuzzy sets, *Journal of Information and Control*, vol.8, pp.338-353, 1965.

[20] C.-Y. Yeh, C.-W. Huang and S.-J. Lee, Multi-kernel support vector clustering for multi-class classification, *International Journal of Innovative Computing, Information and Control*, vol.6, no.5, pp.2245-2262, 2010.

[21] C.-Y. Chen, S.-W. Shyue and C.-J. Chang, Association rule mining for evaluation of regional environments: Case study of Dapeng bay, Taiwan, *International Journal of Innovative Computing, Information and Control*, vol.6, no.8, pp.3425-3436, 2010.

[22] C. C. Chou, A mixed fuzzy expert system and regression model for forecasting the volume of international trade containers, *International Journal of Innovative Computing, Information and Control*, vol.6, no.8, pp.2449-2458, 2010.

[23] T. Tokuyasu, T. Shuto, K. Yufu, S. Kanao, A. Marui and M. Komeda, Fuzzy-based region growing method for detecting thoracic aneurysm from CTA images, *ICIC Express Letters, Part B: Applications*, vol.1, no.2, pp.107-112, 2010.

[24] C. C. Chou, A fuzzy logic approach to dealing with objective data and subjective rating, *International Journal of Innovative Computing, Information and Control*, vol.6, no.5, pp.2199-2210, 2010.

## Appendix.

Symbols used in the article

| Symbol | Description |
|---|---|
| $F$ | A family of fuzzy sets |
| $E$ | Universe of discourse |
| $A$ | A fuzzy set in $E$ |
| $A(x)$ | Membership function of $A$ |
| $A_\alpha = [a_1(\alpha), a_2(\alpha)]$ | An $\alpha$-cut of fuzzy set $A$ |
| $R$ | Real line |
| $(S)$ | Set superimposition operator |
| $(A \cap B)^{(2)}$ | $(A \cap B)$ represented twice |
| $[x_i, y_i]^{(1/n)}$ | Equi-fuzzy intervals with membership $(1/n)$ |
| $X, Y$ | Vectors |
| $\phi_1(x), \phi_2(y)$ | Empirical probability distribution functions for $X$ and $Y$ |
| $E[\phi_1(x)]$ | Expectation of empirical distribution function $\phi_1(x)$ |
| $E[\phi_2(y)]$ | Expectation of empirical distribution function $\phi_2(y)$ |
| $P_1(a, x)$ | Probability distribution function |
| $Var(A)$ | Possibilistic variance of fuzzy number $A$ |
| $S$ | Set of clusters |
| $n$ | Number of input frequent patterns (input clusters) |
| $\theta$ | Threshold |