International Journal of Intelligence Science, 2017, 7, 1-8
http://www.scirp.org/journal/ijis
ISSN Online: 2163-0356
ISSN Print: 2163-0283

A Novel Approach for Clustering Periodic Patterns

Fokrul Alom Mazarbhuiya

Department of IT, College of Computer Science & IT, Albaha University, Albaha, KSA
Email: fokrul_2005@yahoo.com

How to cite this paper: Mazarbhuiya, F.A. (2017) A Novel Approach for Clustering Periodic Patterns. *International Journal of Intelligence Science*, **7**, 1-8.
http://dx.doi.org/10.4236/ijis.2017.71001

Received: October 25, 2016
Accepted: December 27, 2016
Published: December 30, 2016

Abstract

The process of extracting patterns that are frequent from supermarket datasets is a well known problem of data mining. Nowadays, we have many approaches to resolve the problem. Association rule mining is one among them. Supermarket data are usually temporal in nature as they record all the transactions in the supermarket, with the time of occurrence. An algorithm has been proposed to find frequent itemsets, taking the temporal attributes in supermarket dataset. The best part of the algorithm is that each frequent itemset extracted by it is associated with a list of time intervals in which it is frequent. Taking time of transactions as calendar dates, we may get various types of periodic patterns viz. yearly, quarterly, monthly, etc. If the time intervals associated with a periodic itemset are kept in a compact manner, it turns out to be a fuzzy time interval. Clustering of such patterns can be a useful data mining problem. In this paper, we put forward an agglomerative hierarchical clustering algorithm which is able to extracts clusters among such periodic itemsets. Here we take two similarity measures, one on the itemsets of the clusters and others on the corresponding fuzzy time intervals. The efficiency of the proposed method is demonstrated through experimentation on real datasets.

Keywords

Pattern Mining, Temporal Patterns, Locally Frequent Patterns, Superimposition of Intervals, Fuzzy Time-Interval

1. Introduction

The most important data mining problems based on unsupervised learning approach is Clustering and it is very useful for the extraction of data distribution and patterns in the datasets. The clustering process is used to discover both the dense and sparse regions in a dataset. The two main broad approaches are partitioning approach and hierarchical approach. The hierarchical clustering creates a *hierarchy* of clusters from *small to big* or *big to small* and consequently it is named as *agglomerative* or *divisive* clustering techniques respectively. Clustering of numerical data has been studied in the past [1]. However, practically we may have different types of data such as binary, categorical, spatial, ordinal, and temporal or mixture of these. New methods and interesting algorithms for clustering categorical and spatial data have been proposed in the recent times [2] [3] [4] [5].

Association rule mining is an important data mining problem which derives associations among data and was formulated by Agrawal *et al.* [6]. Extracting association rules from *temporal dataset* is also an interesting data-mining problem. In [7], Ale *et al.* have proposed a method of extracting association rules which hold within an itemset's lifetime, where the lifetime of an itemset is the time-period between the first appearance and the last appearance of the itemset in the transactions. In [8], the work proposed by Ale and Rossi [7] is extended by considering time-gap between two consecutive appearance of an itemset in the transactions. The algorithm discussed in [8] mines all locally frequent itemsets along with the list of time-intervals. Each frequent itemsets is associated with a time-intervals lists, where it is frequent. From locally frequent itemsets, we can define various periodic patterns by considering the time-stamps as calendar dates. All such patterns are discussed in detail by the same authors in [9] [10]. While extracting periodicity of a frequent pattern, if the associated time-intervals overlapping, then they can be piled up to form fuzzy interval [11]. Thus, we can have some periodic patterns with each pattern is associated with a fuzzy time interval which describes its period.

In this paper, we devise an agglomerative hierarchical clustering method to explore clusters among such periodic patterns. We define the similarity measure on the corresponding fuzzy time intervals [12] associated with the periodic patterns. Then a *merge* function is defined in terms of the similarity as the union of the pair of periodic frequent itemsets. If the value of the similarity function of fuzzy time intervals is greater than some pre-assigned thresholds, then the corresponding frequent itemsets pairs are *merge* to form larger cluster/itemset and their corresponding fuzzy time intervals are also aggregated [13]. Finally, in this paper we present an algorithm for the clustering of periodic patterns.

The rest of the paper is arranged as follows. Section 2 presents a brief literature review related to the existing clustering algorithms. In section 3, we present some basic definitions and results used in this paper. The proposed agglomerative clustering algorithm is discussed in section 4. In section 5, we discuss some analysis of experiments and results. Finally, we wind up the paper with possible future enhancements of the proposed work in section 6.

2. Related Works

In this segment, we present a brief assessment of the existing research findings related to our work. In [2], *Gibson et al.* have proposed an algorithm for clustering categorical data. Their approach is based dynamical system. In [14], authors have proposed a clustering algorithm called BIRCH. The algorithm proposed in [14] is a hierarchical agglomerative algorithm and it is an efficient algorithm for large datasets. In [4], authors have proposed a robust method for clustering categorical data using summaries. It is known as ROCK and it is an agglomerative hierarchical clustering. In [15], authors have done a survey on clustering time-series data. In [16], authors have discussed algorithm for mining temporal data. In [17], the authors have done a survey on temporal data clustering. Concept of fuzzy sets [18] has been widely used in different areas including cluster analysis, association rule mining, pattern recognition and signal processing in the last couple of years. In [19], Dutta and Mahanta have proposed an algorithm for clustering large categorical database. The approached used in [19] is a fuzzy set based approach. In [20], author has made a survey on fuzzy clustering. In [21], authors have discussed about the applications of fuzzy sets in pattern recognition.

Finding associations among data has also attracted a large number of researchers. In [6], authors have presented a nice and efficient algorithm for association rules extractions. In [7], authors have presented a modified A-priori algorithm for the extractions of temporal association rules. In [1], authors have extended the work of [7] by adding the time-gap between two consecutive transactions containing an itemset. The algorithm [8], gives all locally frequent itemsets along with the lists of time intervals. Here each frequent itemset is linked with a time intervals list in which it is frequent. To compute the periodicity of such frequent itemsets if the time intervals associated with them have overlapping, then a method of redefining the time intervals is proposed in [11], which turns out to be fuzzy time intervals.

3. Problem Definition

In this section, we present a summarized view of some definitions and results on which our proposed algorithm is based.

3.1. Fuzzy Sets

Let X be the universe of discourse, then the fuzzy set A of X is characterized by $A = \{(x, \mu_A(x)); x \in X\}$, where $\mu_A(x)$ = the membership functions of x in A and $\mu_A(x) \in [0,1]$.

3.2. α-Cut of Fuzzy Sets

An α-cut of the fuzzy set A of X is actually a crisp set A_a with elements x of X having membership greater than or equal to a *i.e.*
$A_a = \{(x, \mu_A(x) \geq a, x \in X, a \in [0,1])\}$.

3.3. Convex Fuzzy Sets

A fuzzy set $A = A = \{(x, \mu_A(x)); x \in X\} \subseteq X$ is said to be convex if all it's α-cuts are convex sets.

3.4. Fuzzy Numbers

A fuzzy number is a convex set defined in the real line whose membership value is 1 for at least one $x \in X$.

3.5. Trapezoidal Fuzzy Numbers or Fuzzy Intervals

A trapezoidal fuzzy numbers denoted by $A = (a, b, c, d)$, where it's membership function is given by

$$\mu_A(x) = \left(\begin{array}{c} 0, x \prec a \\ \frac{(x-a)}{(b-a)}, a \le x \le b \\ 1, b \le x \le c \\ \frac{(d-x)}{d-c}, c \le x \le d \\ 0, x \succ d \end{array} \right) \tag{1}$$

In short, we can express the above membership function as

$$\mu_A(x) = \max\left(\min\left(\tfrac{x-a}{b-a}, \tfrac{d-x}{d-c}\right), 0\right) \tag{2}$$

It is to be mentioned here that our fuzzy time intervals associated with periodic frequent patterns are actually trapezoidal fuzzy numbers. The fuzzy time intervals are formed using method [11] and is an L-R fuzzy intervals [22] [23].

3.6. Generalized Trapezoidal Fuzzy Numbers

A generalized trapezoidal fuzzy numbers is represented as $A = (a, b, c, d, h)$, where its membership is given by

$$\mu_A(x) = \left(\begin{array}{c} 0, x \prec a \\ \frac{h(x-a)}{(b-a)}, a \le x \le b \\ h, b \le x \le c \\ \frac{h(d-x)}{d-c}, c \le x \le d \\ 0, x \succ d \end{array} \right) \tag{3}$$

In short, we can express the above membership function as

$$\mu_A(x) = \max\left(\min\left(h\left(\tfrac{x-a}{b-a}\right), h\left(\tfrac{d-x}{d-c}\right)\right), 0\right) \tag{4}$$

3.7. Similarity Measure

Let $0 \le a \le 1$, $0 \le b \le 1$, then the similarity measure between a and b is given by

$$S(a,b) = 1 - \left|\sqrt{a} - \sqrt{b}\right| \tag{5}$$

Let $A = (a1, a2, a3, a4: h_A)$ and $B = (b1, b2, b3, b4: h_B)$ be two generalized trapezoidal fuzzy numbers, then the similarity measure is defined in [12] as follows

$$S(A,B) = \left(\operatorname{sim}(h_A, h_B) \times \prod_{i=1}^{4} \operatorname{sim}(a_i, b_i) \right)^{1/5} \tag{6}$$

The larger the value of $S(A, B)$ the more similarity between A and B. Obviously, $0 \le S(A, B) \le 1$, A and B will be identical if $S(A, B) = 1$.

3.8. Merger of Periodic Frequent Itemsets Belonging to Two Different Clusters

Let A and B be two periodic frequent itemsets having periods T and S respectively. Then the *merge* function is defined $\text{merge}(A, B) = A \cup B$, if and only if $S(T, S) \geq \theta$, where, θ is pre-defined threshold (small positive numbers). It is to be mentioned here that corresponding fuzzy time intervals of T and S associated A and B will be aggregated [13] to form new fuzzy time interval.

4. Proposed Clustering Algorithm

In this segment, we describe our proposed clustering algorithm based on the notion explained in the previous section. The proposed algorithm takes as input, all periodic patterns with fuzzy time intervals describing their periods. The fuzzy time intervals are constructed using the methods discussed in [11]. Assuming that each pattern is associated with exactly one fuzzy time interval, we want to do the clustering of periodic patterns in such a way that each cluster will contain similar type of periodic patterns. The similarity between periodic patterns is defined in terms their corresponding fuzzy time intervals. The similarity measure, $S()$ function is discussed in section III. The itemset A and B having fuzzy time intervals T_1 and T_2 are said to be similar if and only if the value $S(T_1, T_2)$ is greater than some pre-defined threshold.

Initially, each pattern is assigned to a separate cluster. Thereafter, for each pair of clusters the similarity value $S(\ )$ is calculated and merge function is applied (to generate a new bigger cluster) if the $S(\ )$ is greater than the threshold. And their corresponding periods/fuzzy time intervals are aggregated [13]. The process of merging continues till no merger of clusters is possible or there is only one cluster at the top. In bellow we present the pseudo code for the proposed algorithm.

```
Frequent Pattern Clustering Algorithm ($n$, $\theta$)
Input: The number of frequent patterns n and threshold $\theta$
Output: $A$ set of cluster $S$ of with fuzzy time intervals
Steps:
Initially set of clusters is empty
1) $S \leftarrow \phi$
2) read each frequent pattern $A[i]$ with fuzzy time intervals $T[i]$
3)      To construct a cluster $C$ with $T$ if a cluster $C_1 \in S$ with sim($T_1$, $T$) $\geq \theta$
4)          Then $C$ = merge ($C_1$, $C$) with $T$ = aggregate ($T_1$, $T$)
5)            remove $C_1$ and $T_1$ from $S$
6)              add $C$ with its fuzzy time intervals to $S$
7) Process continue till no merger is possible.
8) return $S$
9) stop
```

5. Experiment & Discussions

For experimentation, we have used a synthetic dataset T10I4D100K, available from FIMI[1] website. As the dataset is non-temporal, we consider the temporal features, the calendar dates and execute the algorithm [8] to get the periodic patterns and then execute our clustering algorithm for the threshold value $\theta = 0.4$. The clustering results along with the number of misclassified itemsets obtained are presented in the Table 1. We also represent trend of number of clusters with respect to the number of transactions in graphical form given in Figure 1 and with a bar diagram in Figure 2.

6. Conclusion & Future Works

In this paper, we have presented an agglomerative-hierarchical clustering algorithm to find clusters among periodic patterns with fuzzy time intervals. The

Table 1. Clustering results along with the number of misclassified itemsets for different sets of transactions.

Data Size (No of Transactions)	Max No. of Itemsets	No of Clusters Obtained	Number of Itemsets Misclassified
00000	0	0	0
10000	123	10	3
25000	220	14	2
50000	350	17	1
75000	599	21	1

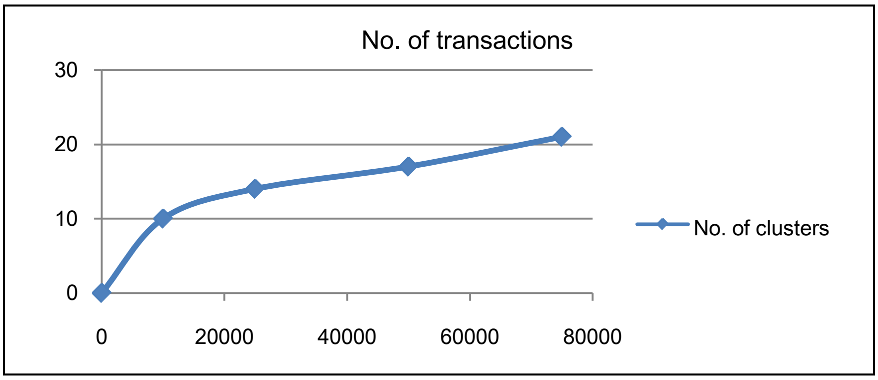

Figure 1. Graph of no. transaction vs. no of clusters/itemsets.

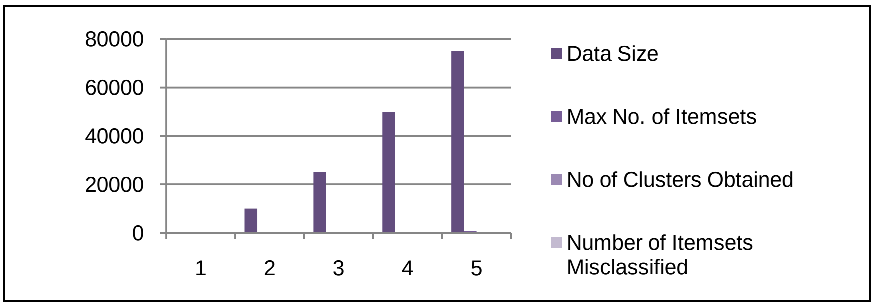

Figure 2. Bar diagram cluster and misclassified items.

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

algorithm starts with as many clusters as the periodic patterns having fuzzy time intervals. Then, if their similarity value is greater than pre-defined threshold, the pairs of clusters are merged. The similarity is defined on fuzzy time intervals associated with periodic patterns. After each level the corresponding fuzzy time intervals are updated by aggregation. Although we have used the agglomerative-hierarchical approach in this paper; any other approach can also be considered provided the similarity measure is properly defined.

References

[1] Hartigan, J.A. (1975) Clustering Algorithms. John Wiley & Sons, Hoboken.

[2] Gibson, D., Kleinberg, J. and Raghavan, P. (1998) Clustering Categorical Data: An Approach Based on Dynamical Systems. In: *Proceedings of the* 24*th International Conference on Very Large Databases*, Morgal Kaufmann, New York, 311-323.

[3] Ng, R.T. and Han, J. (1994) Efficient and Effective Clustering Methods for Spatial Data Mining. In: *Proceedings of the VLDB Conference*, Santiago, 144- 155.

[4] Guha, S., Rastogi, R. and Shim, K. (1999) ROCK: A Robust Clustering Algorithm for Categorical Attributes. In: *Proceedings of the IEEE International Conference on Data Engineering,* IEEE Explore, Sydney, 512-521.

[5] Ganti, V., Gehrke, J. and Ramakrishnan, R. (1999) CACTUS-Clustering Categorical Data Using Summaries. *Proceedings of the* 5*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, 15-18 August 1999 73-83. https://doi.org/10.1145/312129.312201

[6] Agrawal, R., Imielinski, T. and Swami, A.N. (1993) Mining Association Rules between Sets of Items in Large Databases. *Proceedings of the* 1993 *ACM SIGMOD International Conference on Management of Data*, Washington DC, 25-28 May 1993, Vol. 22(2) of *SIGMOD Records*, 207-216. https://doi.org/10.1145/170035.170072

[7] Ale, J.M. and Rossi, G.H. (2000) An Approach to Discovering Temporal Association Rules. *Proceedings of the* 2000 *ACM Symposium on Applied Computing*, **1**, 294-300. https://doi.org/10.1145/335603.335770

[8] Mahanta, A.K., Mazarbhuiya, F.A. and Baruah, H.K. (2005) Finding Locally and Periodically Frequent Sets and Periodic Association Rules. *Proceedings of the* 1*st International Conference on Pattern Recognition and Machine Intelligence*, Kolkata, 20-22 December 2005, 576-582.

[9] Mahanta, A.K., Mazarbhuiya, F.A. and Baruah, H.K. (2008) Finding Calendar-Based Periodic Patterns. *Pattern Recognition Letters*, **29**, 1274-1284.

[10] Mazarbhuiya, F.A. (2014) Discovering Yearly Fuzzy Patterns. *International Journal of Computer Science and Information Security* (*IJCSIS*), **12**, 7-12.

[11] Baruah, H.K. (1999) Set Superimposition and Its Application to the Theory of Fuzzy Sets. *Journal of Assam Science Society*, **10**, 25-31.

[12] Zhou, Y. (2016) A Novel Simiarity Measure for Generalized Trapezoidal Fuzzy Numbers and It's Application to Decision Making. *International Journal of U- and E- Service*, *Science and Technology*, **9**, 131-148. https://doi.org/10.14257/ijunesst.2016.9.3.14

[13] Klir, G.J. and Folger, T.A. (1988) Fuzzy Sets, Uncertainty, and Information. Prentice-Hall of India, New Delhi.

[14] Zhang T., Ramakrishnan, R. and Livny, M. (1996) BIRCH: An Efficient Data Clustering Method for Very Large Databases. 1996 *ACM SIGMOD International Con-*

ference on Management of Data, Canada, 4-6 June 1996, 103-114. https://doi.org/10.1145/233269.233324

[15] Liao, T.W. (2005) Clustering of Time-Series Data—A Survey. *Pattern Recognition*, **38**, 1857-1874. https://doi.org/10.1016/j.patcog.2005.01.025

[16] Rani, Y.L.S., Deepthi, P.W. and Devi, C.R. (2013) Clustering Algorithm for Temporal Data Mining: An Overview. *International Journal of Emerging Technology and Advanced Engineering*, **3**, 350-354.

[17] Yasodha, M. and Ponmuthuramalingam, P. (2012) A Survey on Temporal Data Clustering. *International Journal of Advanced Research in Computer Science and Communication Engineering*, **1**, 768-772.

[18] Zadeh, L.A. (1965) Fuzzy Sets. *Journal of Information and Control,* **8**, 338-353. https://doi.org/10.1016/S0019-9958(65)90241-X

[19] Dutta, M. and Mahanta, A.K. (2004) An Algorithm for Clustering Large Categorical Databases Using a Fuzzy Set Based Approach. 17*th Australian Joint Conference on Artificial Intelligence*, Cairns, 4-6 December 2004, 103-105.

[20] Yang, M.S. (1993) A Survey of Fuzzy Clustering. *Mathematical and Computer Modelling*, **18**, 1-16. https://doi.org/10.1016/0895-7177(93)90202-A

[21] Karim, M.E., Yun, F., Madani, S. and Anderson, E.R. (2010) Fuzzy Clustering Analysis. Master Thesis, Blekinge Institute of Technology, Blekinge.

[22] Dubois, D. and Prade, H. (1983) Ranking Fuzzy Numbers in the Setting of Possibility Theory. *Information Sciences*, **30**, 183-224. https://doi.org/10.1016/0020-0255(83)90025-7

[23] Loeve, M. (1977) Probability Theory. Springer, New York.