ISSN: 2321-7782 (Online)

Volume 4, Issue 1, January 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Clustering of Locally Frequent Patterns over Fuzzy Temporal Datasets using a New Similarity Measure

Fokrul Alom Mazarbhuiya
College of Computer Science & IT
Albaha University,
Kingdom of Saudi Arabia (KSA)

Abstract: Finding patterns from a temporal dataset is a well defined data mining problem. There are many approaches to resolve this problem like association rule mining, clustering and classification. Out these clustering has received a lot of attention among the researchers. Clustering is usually used for discovering data distribution and patterns in a dataset. A couple of algorithms have been proposed so far clustering different types of data. Clustering of fuzzy temporal data is an important extension of temporal data clustering. It is actually the method of finding clusters among the frequent itemsets associated with fuzzy time intervals of frequencies. In this paper, we propose an agglomerative hierarchical clustering algorithm to find clusters among the frequent itemsets obtained from fuzzy temporal data. The efficacy of the proposed method is established through experimentation on a synthetic datasets.

Key words: Clustering, Frequent itemsets, Temporal patterns, Locally frequent itemset, Fuzzy time-interval, aggregate of fuzzy numbers, aggregate of fuzzy intervals, similarity between two itemsets.

I. INTRODUCTION

Clustering is one of the important data mining techniques that follows unsupervised learning approach and it is used to discover data distribution and patterns in the datasets [1]. Association rule mining is another important data mining technique which focuses on deriving associations from data and was formulated by *Agrawal et al* [2]. Mining association rules from *temporal dataset* is an extension of traditional association mining problem. In [3], *Ale et al* proposed an algorithm of extracting association rules which hold througho ut 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. In [4], the work proposed by *Ale and Rossi* [3] is extended by incorporating time-gap between two consecutive transactions containing an itemset. In [5], a method of extracting frequent itemsets from fuzzy temporal data is proposed. The algorithm [5] gives all locally frequent itemsets where each locally frequent itemset is associated with one or more fuzzy time intervals where it is frequent. For the sake of convenience we call locally frequent itemset as frequent itemset. Clustering of such patterns can be interesting. In [6], authors proposed a method clustering such patterns using variance of fuzzy time intervals associated with frequent itemsets.

In this paper, we propose a method to find clusters among frequent patterns/ itemsets based on similarity measure used in [7, 8]. The similarity between two itemsets is defined as ratio of cardinality of their intersection to that of their union.. If the similarity value is less than a pre-assigned threshold, then the corresponding itemsets will be similar and will be merge in a same cluster and their corresponding fuzzy time intervals of frequencies will be aggregated to form a new fuzzy time interval of frequency of the new itemset in the cluster, otherwise they will belong to different clusters. We define a *merge* function in terms of the similarity to merge the i temsetss / clusters.. Thus in each level, each resulting clusters will be associated with a fuzzy time intervals which is aggregate of the fuzzy time intervals associated with frequent itemsets in the previous level..

The paper is organized as follows. In Section-II we discuss about related works. Section-III presents a brief review of the definitions, notations and symbols used in this paper. The proposed algorithm is presented in Section-IV. Section-V gives the experimental results and discussions. Finally, we conclude the paper with possible future enhancements of the proposed work in Section- VI.

II. Related Work

This section presents a brief review of the existing research findings related to our work. An algorithm for clustering categorical data has been proposed in [9]. During the last few years the concept of fuzzy sets [10] has been widely used in different areas including cluster analysis and pattern recognition. In [8], the author proposed an agglomerative algorithm for clustering categorical data using a fuzzy set based approach. Many researchers are attracted to the concept of finding associations among data. A method for the discovery of association rule is presented in [2]. In [3], a method for discovering temporal association rules is described. In [4], the works proposed in [3] is extended by incorporating time-gap between two consecutive transactions containing an item. The algorithm [4] gives all locally frequent itemsets along with the lists of time intervals. An algorithm for finding locally frequent itemsets fro fuzzy temporal data is discussed in [5] where locally frequent itemset is associated with one or more fuzzy time interval. Finding cyclic patterns from such data is discussed in [11]. Clustering of locally frequent itemsets using fuzzy statistical parameter is discussed in [12]. In [13], an agglomerative hierarchical clustering method based on multi view point is discussed. In [14], a method of clustering locally frequent itemsets over fuzzy temporal data using fuzzy variance is discussed.

III. Definition, Notation and Symbols Used

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

Definition 3.1 (*Aggregation operator*)

Let [a_1, b_1, c_1, d_1] and [a_2, b_2, c_2, d_2], ... [a_n, b_n, c_n, d_n] be n fuzzy intervals, the arithmetic mean aggregation operator [15] will produces a fuzzy interval [a, b, c, d] where

$$a = \frac{1}{n}\sum_{1}^{n} a_i \qquad b = \frac{1}{n}\sum_{1}^{n} b_i \qquad c = \frac{1}{n}\sum_{1}^{n} c_i$$

Definition 3.2 (*Similarity of pairs of clusters*).

Let C_1 and C_2 be two clusters and C_1 is associated with fuzzy time intervals *A* which is aggregate of {*A*[i]; i=1,2,…n1} and C_2 is associated with fuzzy time intervals *B* which is aggregate of {*B*[i]; i=1,2,….n2}. The similarity between C_1 and C_2 defined as $sim(C_1, C_2)=| C_1 \cap C_2|/| C_1 \cup C_2|$

Definition 3.3 (*Merger of Clusters)*

Let C_1 and C_2 be two clusters having fuzzy time intervals *A* and *B* 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 U C_2, if and only if $sim(C_1,C_2)\leq\theta$, where θ is a pre-defined threshold value. Here the fuzzy time interval of *C* will be aggregate of *A* and *B*.

IV. Proposed Algorithm

In this section, we present our proposed clustering algorithm based on the concepts discussed in the previous section. For the proposed algorithm, all frequent patterns having fuzzy time intervals describing their interval of frequencies serves as input data. Initially any two frequent itemset will be merged using *merge* () if their similarity defined by *sim*() function is less than some user-specified threshold (merge function and similarity function is defined in section-III). Simultaneously corresponding fuzzy time intervals are merged by a aggregation operator (aggregation operator is defined in section-III) After the first level we will have new itemsets with new fuzzy time intervals. In any level two clusters are merged if their similarity is less than some pre-assigned value along with change in corresponding fuzzy time intervals. So after every level there any change in clusters as well as fuzzy interval. So basically our algorithm is agglomerative hierarchical in nature. The pseudo code of the algorithm is given below.

```
Algorithm Frequent Pattern Clustering (k, θ)

Input:  The number of frequent patterns k and threshold θ.
Output: A set of clusters S
Setps:
start
S ← ϕ
input k, θ
i ← 1
while(i≤ k)
    read  a frequent pattern p[i] with fuzzy time intervals A[i]
    construct a cluster C consisting of p[i] only
    while there is C1 ∈ S having fuzzy time interval A with sim(C1 ,C )≤ θ
            C2 ← merge (C1 ,C) with fuzzy time interval B
            B ← aggregate (A, A[i])
            Remove C1 from S
        C ← C2
        A ←B
    end while
    i ← i+1
    add C to S
end while
return S
stop
```

V. Experimental Setting and Results

For experimental purpose, we have used a synthetic dataset T10I4D100K, available from FIMI1 website. A summarized view of the dataset is presented in Table 1. We incorporate fuzzy time stamp on the dataset to make it suitable for our experiment. We take different sizes of the dataset with different parameters to test the efficacy of our proposed algorithm.

Dataset	# Items	# Transactions	*Min* ǀ Tǀ	*Max* ǀ Tǀ	*Avg* ǀ Tǀ
T10I4D100K	942	100000	4	77	39

Table 1: T10I4D100K dataset characteristics

Dataset No. of transactions	Max no of items	#Clusters obtained	#Itemsets misclassified
10,000	115	7	3
20,000	205	11	4
30.000	253	13	4
40,000	320	15	3
50,000	360	17	3
60,000	478	23	2
100,000	967	25	1

Table 2: Clustering results along with the number of misclassified itemsets for different set of transactions

Thereafter, we have applied the proposed agglomerative-hierarchical algorithm to find clusters from the patterns. For threshold value ($\theta = 0.5$), the clustering results along with the number of misclassified itemsets obtained from the dataset is presented in Table 2, The graphical representation of the results are given in the figure 1 and figure2. It can be observed from Table 2 and Figure2 that with increasing number of transactions in the datasets the number of misclassified items is less.

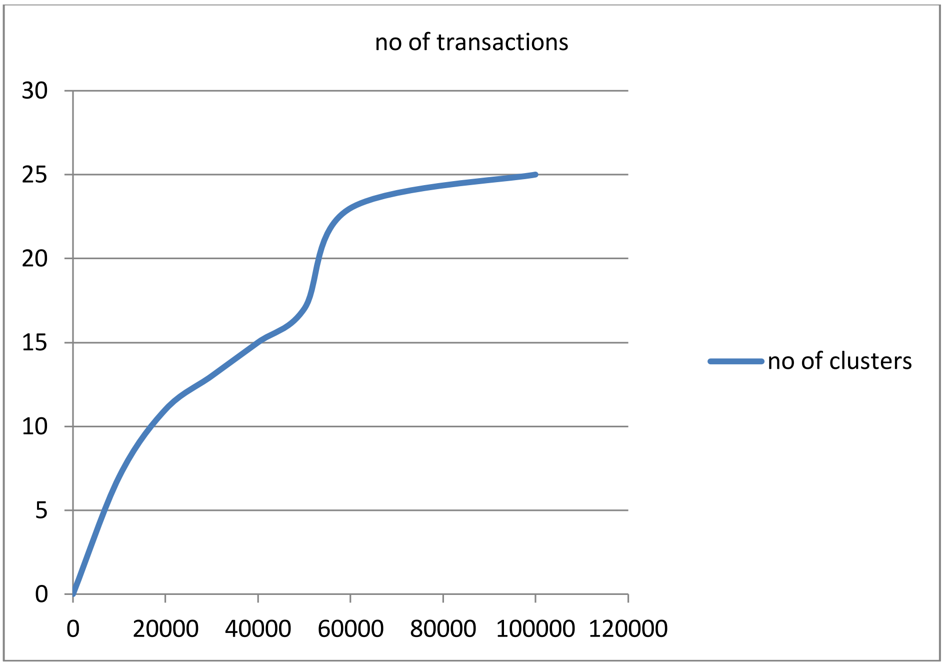


Fig1: no. of transaction vs. no of clusters

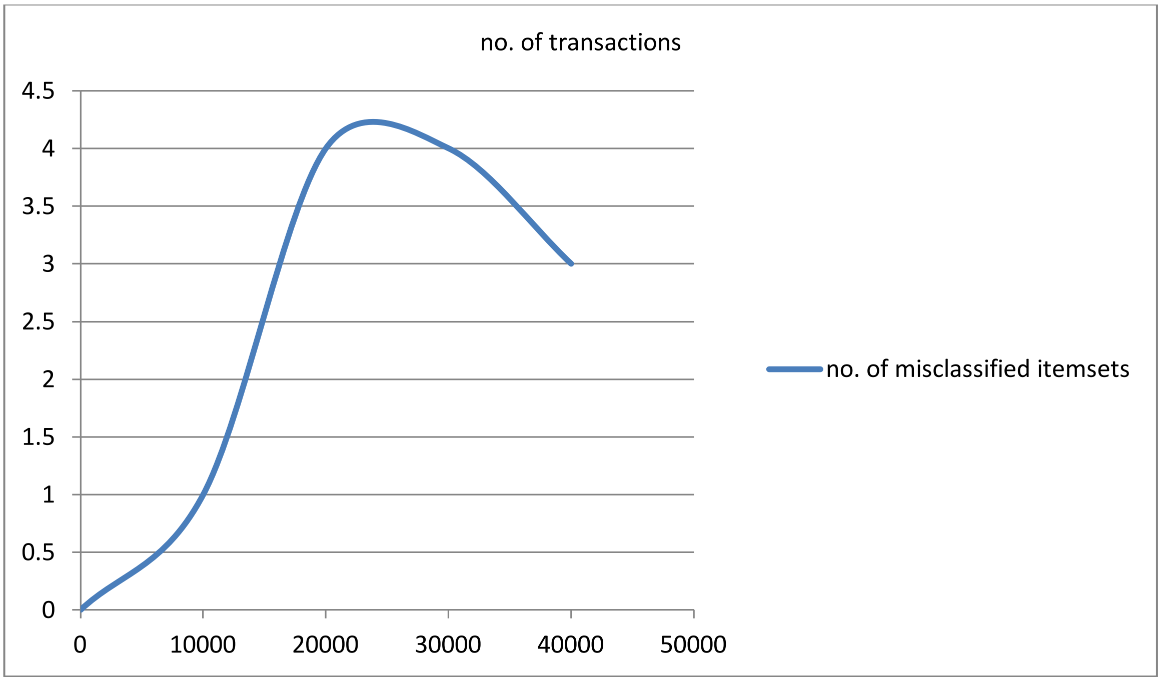


Fig2: no. of transaction vs. no of misclassified itemsets

VI. Conclusions

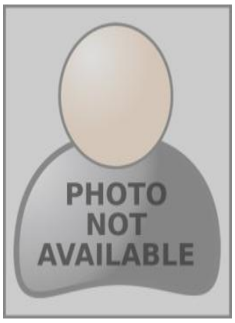
In this paper, we have presented an agglomerative- hierarchical clustering algorithm to find clusters among frequent patterns with fuzzy time intervals. The similarity measure used for clustering is defined in section-III. The algorithm starts with as many clusters as the frequent patterns having fuzzy time intervals. Then, the pairs of clusters are merged if their similarity value is less than a pre-defined threshold and simultaneously their corresponding fuzzy time intervals are merged using aggregation operator given in section-III. 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. .

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 similarity measure and arithmetic mean aggregation operator other measure can be used n future.

References

1. J. A. Hartigan; Clustering Algorithms, John Wiley & Sons, New York, USA, 1975.
2. R. Agrawal, T. Imielinski and A. N. Swami; Mining association rules between sets of items in large databases, In Proc. of 1993 ACM SIGMOD Int'l Conf on Management of Data, Vol. 22(2) of SIGMOD Records, ACM Press, 1993, 207-216.
3. J. M. Ale and G. H. Ross; An approach to discovering temporal association rules, In Proc. of 2000 ACM symposium on Applied Computing, 2000.
4. A. K. Mahanta, F. A. Mazarbhuiya and H. K. Baruah; Finding calendar-based periodic patterns, Pattern Recognition Letters, vol.29, no.9, 2008, 1274-1284.
5. F. A Mazarbhuiya, M. Shenify and Mohammed Husamuddin; Finding Local and Periodic Association Rules from Fuzzy Temporal Data, The 2014 International Conference on Advances in Big Data Analytics, USA, 2014, 240-247.
6. Md. Husamuddin and F. A. Mazarbhuiya; Clustering of Locally Frequent Patterns over Fuzzy Temporal Datasets, International Journal of Computer Trends and Technology (IJCTT), Vol.28 (3), October 2015, India, 131-134.
7. 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.
8. 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, 2001, 103-105.
9. N. K. Sindhu and R. Kaur; Clustering in data mining, International Journal of Computer Trends and Technology (IJCTT), Vol. 4 (4), 2013, 710-714.
10. L. A. Zadeh; Fuzzy Sets , Information and Control Vol. 8, 1965, 338-353.
11. M. Shenify; Extracting Cyclic Frequents Sets from Fuzzy Temporal Data, In proc of the 30th International Conference on Computers and their Applications (CATA-2015), 2015, USA.
12. F. A. Mazarbhuiya, M. Abulaish, Clustering Periodic Patterns using Fuzzy Statistical Parameters. International Journal of Innovative Computing Information and Control (IJICIC), Vol. 8, No. 3(b), 2012, 2113-2124
13. V. V Srivalli, R. G. Kumar, J. Mungara; Hierarchical Clustering With Multi view point Based Similarity Measure, International Journal of Computer Trends and Technology (IJCTT), Vol. 4 (5), 2013, 1475-1480.
14. Md. Husamuddin and F. A. Mazarbhuiya; Clustering of Locally Frequent Patterns over Fuzzy Temporal Datasets, International Journal of Computer Trends and Technology (IJCTT), Vol.28 (3), October 2015, 131-134, .India.
15. George J. Klir and Tina A. Folger, Fuzzy Sets, Uncertainty, and Information, Prentice-Hall of India Pvt. Ltd., New Delhi, 1988

AUTHOR(S) PROFILE



Fokrul Alom Mazarbhuiya received B.Sc. degree in Mathematics from Assam University, India and M.Sc. degree in Mathematics from Aligarh Muslim University, India. After this he obtained the Ph.D. degree in Computer Science from Gauhati University, India. He worked as an Assistant Professor in College of Computer Science, King Khalid University, Saudi Arabia from 2008 to 2011. Curently he is an Assistant Professor in College of Computer Science & IT, Albaha University, Saudi Arabia. His research interest includes Data Mining, Information security, Fuzzy Mathematics and Fuzzy logic.