Clustering of Locally Frequent Patterns over Fuzzy Temporal Datasets

Md Husamuddin[1], Fokrul Alom Mazarbhuiya[2]
[1`]Department of Computer Science, Albaha University, Albaha, Kingdom of Saudi Arabia
[2] College of Computer Science and IT, Albaha University, Albaha, Kingdom of Saudi Arabia

Abstract: *The study of discovering frequent patterns in a dataset is a well defined data mining problem. There are many approaches to resolve this problem including. Clustering is one of the common data mining approaches which is used for discovering data distribution and patterns in a dataset. Many algorithms have been proposed for finding clusters among frequent patterns itemsets. clustering fuzzy temporal data is an extension of temporal data mining. Here we try to find clusters among frequent itemsets based on fuzzy intervals of frequencies. In this paper, we propose a 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 real datasets.*

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

I. INTRODUCTION

Clustering is one of the common data mining problems which follow unsupervised learning approach and it is very useful for the discovering data distribution and patterns in the datasets [1]. Association rule mining is another common data mining problem which focuses on deriving associations among data. The association rule mining problem was formulated by *Agrawal et al* [2]. Mining association rules from *temporal dataset* is also an interesting data-mining problem and recently it has earned a lot of attention. In [3], *Ale et al* have proposed a method of extracting association rules which hold throughout 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 item to solve some of these issues.

The algorithm proposed in [4] outputs all locally frequent itemsets along with the list of time-intervals. In [5], a method of extracting frequent itemsets from fuzzy temporal data is proposed. The algorithm [5] gives as output 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. The fuzzy time interval associated with frequent itemsets exhibits some interesting properties. Both intra itemsets and inter intemsets study can be made. In intra itemset study, the most important example is finding cyclic patterns [6]. In inter itemset study, we derive clustering of the frequent itemsets.

In this paper, we propose a method to find clusters among frequent patterns/ itemsets based fuzzy time intervals associated with them. 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 frequent patterns. Here, we first define similarity between pair of frequent patterns having fuzzy time-intervals as their interval of frequencies 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, we define the similarity between pair of clusters consisting of similar frequent 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, we present an algorithm for the clustering of frequent patterns. The whole approach is similar to that [7] used find clusters from non-fuzzy temporal data.

The paper is organized as follows. In Section-2 we discuss about related works. Section-3 presents a brief review of the definitions, notations and symbols used in this paper. The proposed algorithm is presented in Section-4. Section-5 gives the

experimental results of the proposed algorithm. Finally, we conclude the paper with possible future enhancements of the proposed work in Section- 6.

II. RELATED WORKS

This section presents a brief review of the existing research findings related to our work. In [8], an algorithm for clustering categorical data has been proposed. During the last few years the concept of fuzzy sets [9] has been widely used in different areas including cluster analysis and pattern recognition. In [10], the author has 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. An efficient algorithm for the discovery of association rule is presented in [11]. In [9], an algorithm for discovery of temporal association rules is described. In [10], the works proposed in [9] is extended by incorporating time-gap between two consecutive transactions containing an item. The algorithm [10] 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 [6]. Clustering is one of the important data mining techniques [12] and is applied to different types of datasets ([13], [14]). In [15], an agglomerative hierarchical clustering method based on multi view point is discussed. 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.

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 *(Possibilistic mean and possibilistic variance of a fuzzy number).*

Let F be a family of fuzzy number and A be a fuzzy number belonging to F. Let $A_\alpha=[a_1(\alpha), a_2(\alpha)],\alpha\in[0, 1]$ be an α-cut of A. The interval-valued possibilistic mean [16] of fuzzy number $A\in F$ is defined as

$$M(A) = [\, M_*(A), M^*(A)\,]$$

where the lower possibilistic mean value of A is expressed as

$$M_*(A) = 2\int_0^1 \alpha a_1(\alpha)d\alpha$$

Similarly, the upper possibilistic mean value of A is expressed as

$$M^*(A) = 2\int_0^1 \alpha a_2(\alpha)d\alpha$$

And the possibilistic variance [16] of $A\in F$ is expressed as

$$\text{Var}(A) = \tfrac{1}{2}\int_0^1 \alpha(a_2(\alpha)-a_1(\alpha))^2 d\alpha$$

Thus the variance of A is the expected value of the squared-deviations between arithmetic mean and the endpoints of its α-cuts. Also the variance of a fuzzy number is invariant to shifting. [16].

Definition 3.2 (Similarity measure between pairs of frequent pairs of frequent patterns having fuzzy time intervals as their interval of frequencies). Let A_1 and A_2 be two frequent patterns with fuzzy time intervals T_1 and T_2 respectively. The similarity measure [7] between A_1 and A_2 is represented as $sim(A_1,A_2)$ and defined in the equation 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.

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

We consider two patterns as similar if and only if value of their *sim function* [7] 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,…n1} and C_2 consists of similar patterns say {B[i]; i=1,2,….n2}. The similarity between C_1 and C_2 is defined using equation in which $D_1= \sum_{i=1}^{n1} var(T[1])/n_1$ is the average of the variances of {T[i]; I = 1,2,……,n1} that are associated with the similar patterns {A[i]; i= 1,2,….,n1} of C_1 and $D_2= \sum_{i=1}^{n2} var(T[1])/n_2$ is the average of variances of {T[i]; I = 1,2,……,n1} that are associated with similar patterns {B[i]; I = 1,2,….,n2} of C_2.

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

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 U C_2, if and only if sim(C_1,C_2)$\leq\theta$, where θ is a pre-defined threshold value [see e.g. [7]]

IV. PROPOSED ALGORITHM

In this section, we present the proposed clustering algorithm based on the concepts discussed in the previous section [7]. The algorithm is similar to the algorithm [7] that is used to find clusters among periodic patterns [4]. The dataset associated with [7] is temporal however in our case it is fuzzy temporal. For the proposed algorithm, all frequent patterns having fuzzy time intervals describing their interval of frequencies serves as input data. The methods to find frequent patterns having fuzzy time intervals as their time intervals of frequencies are discussed in [5] which is termed as locally frequent itemsets over fuzzy time intervals. Since the variance of fuzzy intervals are invariant to shifting, two periodic patterns having the same value of variance for the fuzzy time intervals describing their intervals of frequencies can be considered to be similar. Considering frequent patterns along with fuzzy time intervals describing their interval of frequencies (each pattern is associated with one or more 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 section-3) 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 values is calculated and merged 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. The pseudo code for the proposed 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:

1. start
2. $S \leftarrow \phi$
3. input k, θ
4. $i \leftarrow 1$
5. while($i \leq k$)
6. read a frequent pattern p[i]
7. construct a cluster C consisting of p[i] only
8. while there is $C_1 \in S$ with sim(C_1,C) $\leq \theta$
9. $C_2 \leftarrow$ merge (C_1 ,C)
10. Remove C_1 from S
11. $C \leftarrow C_2$
12. end while
13. $i \leftarrow i+1$
14. add C to S
15. end while
16. return S
17. 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.

Table 1: T10I4D100K dataset characteristics

Dataset	# Items	# Transactions	*Min* | T|	*Max* | T |	*Avg* | T|
T10I4D100K	942	100000	4	77	39

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

Dataset	Max no of items	#Clusters obtained	#Itemsets misclassified
T1	115	8	4
T2	205	10	4
T3	253	12	3
T4	320	14	3
T5	360	18	2
T6	478	22	2
T7	967	25	1

Thereafter, we have applied the proposed agglomerative-hierarchical algorithm to find clusters among the patterns. For threshold value (θ = 0.4), the clustering results along with the number of misclassified itemsets obtained from the dataset is presented in Table 2. It can be observed from Table 2

that with increasing number of transactions in the datasets the number of misclassified items is less.

VI. CONCLUSION

In this paper, we have presented an agglomerative-hierarchical clustering algorithm to find clusters among frequent patterns with fuzzy time intervals. The algorithm is similar to [7] that is used to find clusters among periodic patterns discussed in [4]. 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. 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.

REFERENCES

[1] J. A. Hartigan (1975);*Clustering Algorithms*, John Wiley & Sons, New York, USA.

[2] R. Agrawal, T. Imielinski and A. N. Swami (1993), 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, pp 207-216.

[3] J. M. Ale and G. H. Rossi (2000); An approach to discovering temporal association rules, *In Proc. of 2000 ACM symposium on Applied Computing*.

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

[5] F. A Mazarbhuiya, M. Shenify and Mohammed Husamuddin (2014); Finding Local and Periodic Association Rules from Fuzzy Temporal Data, *The 2014 International Conference on Advances in Big Data Analytics*, USA.

[6] M. Shenify, (2015); Extracting Cyclic Frequents Sets from Fuzzy Temporal Data, *In proc of the 30th International Conference on Computers and their Applications (CATA-2015)*, USA.

[7] F. A Mazharbhuiya and Muhammad Abulaish (2012); Clustering periodic frequent patterns using fuzzy statistical parameters, *International journal of innovative computing, Information and control*, vol.8, no.3(B), pp.2113-2124.

[8] M. Dutta, A. K. Mahanta and M. Mazumder (2001); 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.

[9] L. A. Zadeh (1965); Fuzzy Sets, *Information and Control* Vol. 8, pp. 338-353.

[10] M. Dutta and A. K. Mahanta (2004); 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.

[11] R. Agrawal and R. Srikant (1994); Fast Algorithms for Mining Association Rules, *In Proc. of the 20th VLDB Conf.*, Santiago, Chile, 1994.

[12] N. K. Sindhu and R. Kaur (2013); Clustering in data mining, International Journal of Computer Trends and Technology (IJCTT), Vol. 4 (4), pp.710-714.

[13] A. N. Sravya, and M. Nalini Sri (2013); A vovel approach of temporal data clustering via weighted clustering ensemble with different representation, International Journal of Computer Trends and Technology (IJCTT), Vol. 4 (4), pp. 624-629.

[14] P. P. Pradhan, D. Mishra, S. Mishra, and S. Shaw (2013); Artificial Bee based Optimized Fuzzy c-Means Clustering of Gene Expression Data, International Journal of Computer Trends and Technology (IJCTT), Vol. 4 (5), pp 1-5.

[15] V. V Srivalli, R. G. Kumar, J. Mungara (2013); Hierarchical Clustering With Multi view point Based Similarity Measure, International Journal of Computer Trends and Technology (IJCTT), Vol. 4 (5), pp. 1475-1480.

[16] C. Carlsson and R. Fuller (2001); On Possibilistic Mean Value and Variance of Fuzzy Numbers, *Fuzzy Sets and Systems* 122 (2001), pp. 315-326.