The Journal of
Mathematics and Computer Science

Available online at
http://www.TJMCS.com

The Journal of Mathematics and Computer Science Vol .2 No.1 (2011) 37-43

Rule Extraction for Blood Donators with Fuzzy Sequential Pattern Mining

Fatemeh Zabihi[1], Mojtaba Ramezan[2], Mir Mohsen Pedram[3], Azizollah Memariani[4]

Industrial Engineering Department, Faculty of Engineering, Shomal University, Amol, Iran, fatemehzabihy@yahoo.com
Computer Engineering Department, Tarbiat Moallem University, Karaj/Tehran, Iran, mojtaba.ramezan@gmail.com
Computer Engineering Department, Faculty of Engineering, Tarbiat Moallem University, Karaj/Tehran, Iran, pedram@tmu.ac.ir
School of Economic Sciences, Tehran, Iran, memar@irphe.ir

Received: August 2010, Revised: November 2010
Online Publication: January 2011

Abstract

Sequential pattern mining is to discover all sub-sequences that are frequent. The classical sequential pattern mining algorithms do not allow processing of numerical data and require converting these data into a binary representation, which necessarily leads to a loss of information. Fuzzy sets are used to overcome this problem and fuzzy set based algorithms have been proposed to handle numerical data using intervals, particularly fuzzy intervals. In this paper, a fuzzy sequential pattern mining algorithm is applied to mine fuzzy sequential patterns from the Blood

[1]MSc in Industrial Engineering
[2]BS in Computer Engineering
[3]Assistant Professor of Electrical & Computer Engineering
[4]Professor of Operation Research (OR)

Transfusion Service Center data set. It helps to predict future patterns of blood donating behavior.

Keywords: Sequence, Fuzzy Sequential Pattern Mining, Fuzzy rules.

1. Introduction

Data mining is able to extract interesting pattern from databases and the bottleneck of knowledge acquisition can be eased in building expert systems [1].Knowledge acquisition can be easily achieved for users by checking the patterns discovered from databases. An important type of knowledge representation is association rule. A method to find the association rules has initially proposed by Agrawal and et al [2]; they also proposed the Apriori algorithm later [3]. However, another important type of knowledge representation is sequential pattern mining. Discovering sequential patterns from large sequence databases are an important problem in the field of knowledge discovery and data mining. In recent years, sequential pattern mining has broadly applied to several application domains, such as market-basket data analysis, medicine, Web log analysis, and telecommunications, etc. Some algorithms for mining sequential patterns were proposed by Agrawal and Srikant [4]. A sequence is an ordered list of itemsets [4].

In real-life application, traditional sequential pattern mining approach may bear the crisp boundary problem. For example, suppose a crisp boundary of minimum support, if sequence support is close to this boundary of minimum support, the sequence is either pruned or selected as a sequential pattern; while fuzzy sets allow partial membership in sets in compare to crisp sets in which each element is either completely in a set or not. Thus a powerful mechanism for representing vague concepts is provided. In previous study, a number of researchers have exploited fuzzy techniques to mine fuzzy association rules [5]-[8] or sequential patterns from databases [6], [9], [10], such as fuzzy support and confidence measure [6], [7], fuzzy quantitative mining [5], [8], and fuzzy time-interval mining [9].

Also Zabihi and et al [11] proposed a novel fuzzy sequential pattern mining algorithm which searched sequences in databases with sliding window constraint that permits elements of a pattern to span a set of transactions within a user-specified window. So loss of useful sequences is prevented in the search process. In addition, Zabihi worked on fuzzy sequential pattern mining with sliding window and time gap constraint in [12]. She proposed six algorithms which will be able to process quantitative data but also prevent from deleting efficient sequences that occurred in crisp sequential pattern algorithms and generate relevant patterns based on the defined time constraints.

In this paper, a fuzzy sequential pattern algorithm is applied to mine fuzzy sequential patterns from the Blood Transfusion Service Center data set. This algorithm finds the relation between blood donating times and donating in a special month. For the number of donating a fuzzy membership function is used. According to this the probably of donating in a specific month will be computed. It finds donating frequency for every donator and helps to predict future patterns of blood donating behavior.

1. Sequential pattern mining

Sequential Pattern mining refers the problem of discovering the existent maximal frequent sequences in a given database. The problem was first introduced by Agrawal and Srikant [13], [14], in which pattern detection was the basic concept. It seeks similar patterns in data

transaction; this approach is useful when the data to be mined has some sequential nature to deal with databases that have time-series characteristics, i.e. each piece of data is an ordered set of elements [15]. For example, it can be said that 60% of people who donate their blood (X) will donate their blood (Y) afterward (X⇒ Y), regardless of the time gap. Sequential Pattern can be defined as follows.

Definition 1: Let I = {$x_1..x_n$} be a set of items. An itemset is a non-empty subset of items, and an itemset with k items is called k-itemset. A sequence s = ($X_1..X_m$) is an ordered list of itemsets. In a set of sequences, a sequence s is maximal if s is not contained in other sequences [16].

All records from an object are grouped together and sorted in increasing order of their timestamp. They are called a data sequence. An object supports a sequence s if it is included within its data sequence (s is a subsequence of the data sequence). The support of a sequence (supp(s)) is defined as the percentage of objects supporting s. In order to decide whether a sequence is frequent or not, a minimum support value (minSupp) is specified by the user. The sequence is said to be frequent if the condition supp(s) ≥ minSupp holds. Given a database of object records, the problem of sequential pattern mining is to find all maximal frequent sequences [13]. Note that items are processed using a binary evaluation – present or not present. In our case, we use fuzzy membership function for frequency of blood donating and it has driven us to fuzzy intervals and so to fuzzy sequential patterns mining.

3. Fuzzy Sequential Patterns

In order to mine fuzzy sequential patterns, the universe of each quantitative item is partitioned into several fuzzy sets. The attribute and itemset concepts have been redefined relative to classical sequential patterns, as in [17].

Definition 2: A fuzzy item is the association of one item and one corresponding fuzzy set. It is denoted by [x, a] where x is the item (also called attribute) and a is the associated fuzzy set.

Example 1: [frequency, short] is a fuzzy item where short is a fuzzy set defined by a membership function on the quantity universe of the possible values of the item frequency.

Definition 3: A fuzzy itemset is a set of fuzzy items. It can be denoted as a pair of sets (set of items, set of fuzzy sets associated to each item) or as a list of fuzzy items.

We will note: (X, A) = ([x_1, a_1], ..., [x_p, a_p]), where X is aset of items, A and [x_i, a_i] are a set of corresponding fuzzy sets and fuzzy items, respectively.

Example2: (X, A) = ([pen, lot][butter, little]) is a fuzzy itemset.

One fuzzy itemset contains only one fuzzy item related to one single attribute. For example, the fuzzy itemset ([pen, lot][pen, little]) is not a valid fuzzy itemset because it contains twice the attribute length. Lastly we define a g-k-sequence.

Definition 4: A g-k-sequence S =< $s_1 \cdots s_g$> is a sequence constituted by g fuzzy itemsets s = (X,A) grouping together k fuzzy items [x, a].

Example 3: The sequence < ([pen, lot][butter, little]) ([cheese, lot]) > groups together 3 fuzzy items into 2 itemsets. It is a fuzzy 2-3-sequence.

4. Algorithm Steps

Association rules are statements of the form {X_1, X_2, …, X_n}⇒ Y, meaning that if all of X_1, X_2,…, X_n are found in the database, then there is a good chance of finding Y. The probability of finding Y

for us to accept this rule is called the confidence of the rule. We normally would search only for rules that had confidence above a certain threshold.

In this algorithm, rules are generated and extracted from database. Because the extracted rules happen via time sequence, they are considered as sequential patterns. So there is no need to compute confidence.

Nomenclature:

- *ItemSet*: one or more items with the same time-stamp,
- *gS*(goal sequence): one or more *ItemSets* to be qualified by the proposed algorithm,
- *Object*: each donator,
- *curIS*: current *ItemSet*,
- *R*: fuzzified attribute for the current *Object*,
- *status*: a flag that indicates whether the current *Object* supports the *gS* or not,
- *count*: a counter to compute the support of *gS*

The proposed algorithm was run on the Blood Transfusion Service Center data setas follow:

- Step 1: *gS*←Generate a goal-sequence
- Step 2: **If** *gS*≠∅ **Then**
    - **For each** *Object* **Do**
        - **put** the first *itemSet* of *gS* in *curIS*,
        - **clear** *status*,
        - **For each** record of *Object* (*R*) look for items of *curIS*, **If** all of its items have been found and its membership degree (as shown in figure 1) are greater than ω, then **set** *status* and **break**,
        - **If***status*=**set**, then increment *count*,
    - calculate the support of *gS*, (that is the ratio of *count* to the number of *Objects*)
    - **Go to** Step1.
- Step 3: **End**

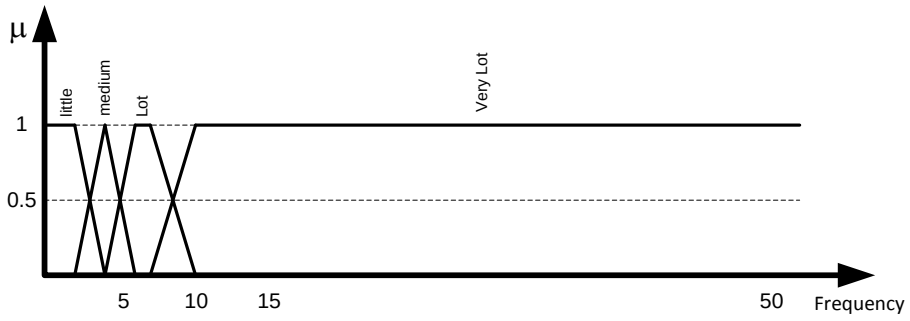

Figure 1: Fuzzy membership functions

5. Experiment result

Given a Blood Transfusion Service Center data set, to run our algorithm, we selected 748 donors at random from the donor database as training set. Each record of the dataset includes the following attributes: R (Recency - months since last donation), F (Frequency- total number of

donation), M (Monetary-total blood donated in c.c.), T (Time-months since first donation), and a binary variable representing whether he/she donated blood in March 2007 (1 stand for donating blood; 0 stands for not donating blood). Table 1 shows 6 records of 748 donors (records).

Table 1: Blood Transfusion Service Center Data Set

Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	whether he/she donated blood in March 2007
2	1	250	2	1
11	1	250	11	1
21	3	750	40	0
3	5	1250	12	1
1	9	2250	51	0
1	13	3250	47	0

80% of records of the database have randomly been chosen as training set and the algorithm has been run for the training set, and rules were extracted with their support as shown in table 2, 3, and 4. Each table shows one factor, i.e. Frequency, Recency, and Time. For support of extracted rules a validity threshold is considered and whether or not the support of rules is greater than this threshold, the reliable rules are attained. For the rest of the records the precision of extracted rules was computed.

Table 2: Result of running the fuzzy sequential pattern algorithm (Frequency)

Fuzzy rules	Extracted fuzzy sequential pattern	Support (80%)
Frequency= Few ? Donation= No	[Frequency, Few][Donation, No]	0.4183
Frequency= Few ? Donation= Yes	[Frequency, Few][Donation, Yes]	0.0633
Frequency= Medium ? Donation= No	[Frequency, Medium][Donation, No]	0.225
Frequency= Medium ? Donation= Yes	[Frequency, Medium][Donation, Yes]	0.0633
Frequency= Lot ? Donation= No	[Frequency, Lot][Donation, No]	0.2083
Frequency= Lot ? Donation= Yes	[Frequency, Lot][Donation, Yes]	0.0867
Frequency= Very Lot ? Donation= No	[Frequency, Very Lot][Donation, No]	0.1333
Frequency= Very Lot ? Donation= Yes	[Frequency: Very Lot, Donation: Yes]	0.0833

Table 3: Result of running the fuzzy sequential pattern algorithm (Recency)

Fuzzy rules	Extracted fuzzy sequential pattern	Support (80%)
Recency= Few ? Donation= No	[Recency, Few][Donation, No]	0.1683
Recency= Few ? Donation= Yes	[Recency, Few][Donation, Yes]	0.0933
Recency= Medium ? Donation= No	[Recency, Medium][Donation, No]	0.145
Recency= Medium ? Donation= Yes	[Recency, Medium][Donation, Yes]	0.0833
Recency= Lot ? Donation= No	[Recency, Lot][Donation, No]	0.045
Recency= Lot ? Donation= Yes	[Recency, Lot][Donation, Yes]	0.0117
Recency= Very Lot ? Donation= No	[Recency, Very Lot][Donation, No]	0.455

Recency= Very Lot ? Donation= Yes	[Recency: Very Lot, Donation: Yes]	0.0533

Table 4: Result of running the fuzzy sequential pattern algorithm (Time)

Fuzzy rules	Extracted fuzzy sequential pattern	Support (80%)
Time= Few ? Donation= No	[Time, Few][Donation, No]	0.1417
Time= Few ? Donation= Yes	[Time, Few][Donation, Yes]	0.0417
Time= Medium ? Donation= No	[Time, Medium][Donation, No]	0.3033
Time= Medium ? Donation= Yes	[Time, Medium][Donation, Yes]	0.0983
Time= Lot ? Donation= No	[Time, Lot][Donation, No]	0.2317
Time= Lot ? Donation= Yes	[Time, Lot][Donation, Yes]	0.09
Time= Very Lot ? Donation= No	[Time, Very Lot][Donation, No]	0.2483
Time= Very Lot ? Donation= Yes	[Time: Very Lot, Donation: Yes]	0.0683

The highlighted records are the rules satisfying the considered validity threshold as shown in table 2, 3, and 4. For these rules, precision is calculated as shown in table 5. The extracted rules show the support of blood donation is not reliable as dominant rules in the considered date. In fact, the obtained result is for the sake of limited number of rules which support item [donation, yes].

Table 5: computed precision

	Frequency	Recency	Time
precision	88%	72%	94%

6. Conclusion

In this paper, we used fuzzy sequential pattern algorithm to extract rules in Blood Transfusion Service Center data set. The rules can be applied for prediction of behavior of donator in the future. For example, how often a donation is done, depends on the donator's previous behavior. It helps service center to schedule and predict the blood supply to ensure an adequate stock of blood for future needs such as accident victims, surgeries, and people suffering from certain diseases, as well as for medical research.

References

[1] Hong, T.P., Wang, T.T., Wang, S.L., and Chien, B. C., "Learning a coverage set of maximally general fuzzy rules by rough sets", *Expert Systems with Applications*, pp. 97-103, 2000.

[2] Agrawal, R., Imielinski, T., & Swami, A., "Mining association rules between sets of items in large databases", *Proceeding of the ACM SIGMOD International Conference on Management of Data*, pp. 207-216, 1993.

[3] Agrawal, R., Mannila, H., Srikant, Toivonen, R. H., and Verkamo, A. I., "Fast discovery of association rules",*In Fayyad U. M., Piatetsky-Shapiro G., Smyth P., and Uthurusamy R., Advances in knowledge discovery and data mining,* AAAI Press, 1995.
[4] Agrawal, R., and Srikant, R., "Mining sequential patterns", *Proseedings of the Eleventh International Conference on Data Engineering*, pp. 3-14, 1995.
[5] Wang S. L., Kuo C. Y., and Hong T. P., "Mining fuzzy similar sequential patterns from quantitative data", *IEEE International Conference on Systems, Man and Cybernetics, Hammamet, Tunisia*, 2002.
[6] Luo J. and Bridges S. M., "Mining fuzzy association rules and fuzzy frequency episodes for intrusion detection", *International Journal of Intelligent Systems*, Vol. 15, No. 8, pp. 687-703, 2000.
[7] Kuok C. M., Fu A., Wong M. H., "Mining fuzzy association rules in databases", *SIGMOD Record*, Vol. 27, No. 1, pp.41-46, 1998.
[8] Zhang W., "Mining fuzzy quantitative association rules", *Proceedings 11th International Conference Tools Artificial Intelligence, Chicago, IL*,pp. 99-102, 1999.
[9] Chen Y. L. and Huang C. K., "Discovering fuzzy time-interval sequential patterns in sequence databases", *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 35, No. 5, pp. 959-972, 2005.
[10] Chen R. S., Tzeng G. H., Chen C. C., and Hu Y. C., "Discovery of fuzzy sequential patterns for fuzzy partitions in quantitative attributes", *ACS/IEEE International Conference on Computer Systems and Applications*, pp. 144-150, 2001.
[11] Zabihi F., Ramezan M., Pedram M.M., and Memariani A., "Fuzzy Sequential Pattern Mining with Sliding Window Constraint", *2nd International Conference on Education Technology and Computer (ICETC), Shanghai, China*, 2010.
[12] Zabihi F., Supervisor: Pedram M.M. and Advisor: Memariani A., "Fuzzy Constrained Sequential Pattern Mining", *A dissertation submitted in partial fulfillment of the requirements for the degree of Master of Industrial Engineering (Industrial Engineering) in the Tarbiat Moallem University, Tehran, Iran* ,2008.
[13] Agrawal R. and Srikant R. "Mining Sequential Patterns", *IBMAlmaden Research Center, 650 Harry Road, San Jose, CA 95120-6099,* 2009.
[14] Agrawal R. and SrikantR., "Mining Sequential Patterns: Generalizations and Performance Improvements", *IBMAlmaden Research Center, 650 Harry Road, San Jose*, CA 95120,1996.
[15] Antunes C. and Oliveira A., "Sequential Pattern Mining Algorithms: Trade-offs between Speed and Memory", *Instituto Superior Tcnico /INESC-ID*.
[16] Kaya M. Alhajj R., "Multi-Objective Genetic Algorithm Based Approach for Optimizing Fuzzy Sequential Patterns", *16th IEEE International Conference on Tools with Artificial Intelligence*, 1082- 3409/04, 2004
[17] Fiot C., Laurent A., and Teisseire M., "Motifs séquentielsflous: un peu, beaucoup, passionnément", *in 5èmes journéesd' Extraction et Gestion des Connaissances*, pp. 507–518, 2005.