A Multi-Criteria Decision Making Based Method for Ranking Sequential Patterns

Zeinab Dashti, Mir Mohsen Pedram, Jamshid Shanbehzadeh

Abstract— **Sequences are one of the most important types of data. Recently, mining and analysis of sequence data has been studied in several fields. Sequence database mining and change mining is an example of data mining to study temporal data. Specific changes might be important to decision maker in different time periods to schedule future activities. Working with long sequences requires useful method. This paper presents a study on similarity measure and ranking sequence data. We employed sequence distance function based on structural features to measure the similarity, and a multi-criteria decision making techniques to rank them.**

Index Terms— **Sequences similarity; distance function; conditional probability distribution; multi-criteria decision making; TOPSIS.**

I. Introduction

In recent decade, significant evolutions are developed in data mining techniques. These techniques are applied in various and successful applications in different domains, e.g. marketing, investment and banking. One of the important tasks of data mining is database mining and change mining in temporal databases. It has been seen in such cases that some patterns exist in one time period but change in other period [3]. For example, "Computer.Memory.Color_Printer" can be thought as a sequence used frequently in some year, but it changed into "Computer.Memory.Multifunctional_Printer" in the next year.

Change mining is an example of data mining for studying time-varying data which tends to discover, analyze and interpret changes. It includes methods that capture changes and analyzes the current and future changes. Discovering and tracking the pattern changes helps decision makers for better decision making. Several studies have been done to detect and describe the difference between two sets of patterns, but they've been limited to association rules. A recent research was reported on changes of sequence patterns in two different periods which determines the changes based on patterns distance [11], while it does not give information about change direction. When there is a need to compare and rank some sequences or patterns extracted from different periods, there should be a distance measure as well as ranking method. In such cases, the decision maker's criteria will play an essential role in determining how well a pattern is satisfying.

In this paper a new method is proposed to compare sequence patterns and rank them based on multi-criteria decision making.

The rest of this paper is organized as follows. In Section 2, we review multi-criteria decision making. Then, TOPSIS method is described in Section 3. Section 4 discusses about distance measures for sequences. A new method is introduced to compare and rank different sequences in Section 5. Finally, a brief conclusion is given in Section 6.

II. Multi-criteria Decision Making

Decision making is a part of our daily lives. In decision science, decision making problems are classified into the following categories [6]: (1) multi attribute decision making (MADM), (2) multi objective decision making (MODM). The major difference of the two classes is in existence of predetermined alternatives. MODM deals with optimization problems in which several objective functions should be satisfied, while MADM is associated with the problems in which alternatives have been predetermined. It means making preference decisions (e.g., evaluation, prioritization, selection) over the available alternatives that are characterized by multiple, usually conflicting, attributes [13]. MADM methods are widely used for real world problems [7, 4, 10].

III. TOPSIS

Researchers have proposed several methods for MADM problems, such as ELECTRE [9] and TOPSIS [5]. The methods cannot be used in a case that ideal alternatives and weights of criteria are unknown. One of applicable methods in such cases is LINMAP method [8, 15]. LINMAP and TOPSIS are different in the types of information that they need [2].

In this paper, TOPSIS is used to rank alternatives. In the TOPSIS method, decision making matrix and weight vector are determined as crisp values and a positive ideal solution (PIS) and a negative ideal solution (NIS) are obtained from the decision matrix. TOPSIS is based on the idea that the proper alternative has the shortest distance from PIS and the longest distance from NIS. TOPSIS ranks the alternatives according to these two distance measures. Decision matrix is

Manuscript received January 12, 2010.

Z. Dashti is MS student in the Computer Engineering Department at Islamic Azad University Science and Research Branch, Tehran, Iran (e-mail: ze.dashti@yahoo.com).

M. M. Pedram, is with the Computer Engineering Department, Faculty of Engineering, Tarbiat Moallem University, Karaj/Tehran, Iran (e-mail: pedram@tmu.ac.ir).

J. Shanbehzadeh is with the Computer Engineering Department, Faculty of Engineering, Tarbiat Moallem University, Karaj/Tehran, Iran (e-mail: shanbehzadeh@gmail.com).

often employed in MADM to start the evaluation process [8] and the evaluation of alternatives A_1, A_2, …, A_n are performed according to criteria B_1, B_2, …, B_m. Criteria might have different dimensions. For a simpler comparison and evaluation, based on all the criteria in a dimensionless units, values are normalized [1, 5] which also help to avoid computational complexity, resulting from different measures in decision matrix.

The steps of TOPSIS method are as follow:

- First step: Convert decision matrix with m alternatives and n criteria to a dimensionless matrix (x_{ij} is the value of *i*th alternative in jth criteria),

$$r_{ij} = x_{ij} \Big/ \left(\sum_{i=1}^{m} x_{ij}^2 \right)^{\frac{1}{2}} \quad , \quad i = 1, \dots, m\,;\ j = 1, \dots, n \qquad (1)$$

- Second step: Obtain a weighted normalized decision matrix,

$$v_{ij} = w_j\, r_{ij} \qquad , \quad i = 1, \dots, m\,; j = 1, \dots, n \qquad (2)$$

where w_j is the weight of *j*th criteria.

- Third step: Determine the positive ideal solution (A^*) and negative ideal solution (A^-).

$$A^* = (\, v_1^*, \dots, v_j^*, \dots, v_n^* \,) = \left\{ \begin{matrix} (\max_j v_{ij} \ \mid j \in J), \\ (\min_j v_{ij} \ \mid j \in J') \mid i = 1, \dots, m \end{matrix} \right\} \qquad (3)$$

$$A^- = (\, v_1^-, \dots, v_j^-, \dots, v_n^- \,) = \left\{ \begin{matrix} (\min_j v_{ij} \ \mid j \in J'), \\ ((\max_j v_{ij} \ \mid j \in J) \mid i = 1, \dots, m \end{matrix} \right\} \qquad (4)$$

V_{j*} and V_{j-} are the best and the worst weighted normalized values for all alternatives according to *j*th criterion, respectively. J is the set of benefit attributes while J' is the set of cost attributes.

- Fourth step: Calculate Euclidean distance from *i*th alternative to positive ideal solution and negative ideal solution.

$$S_i^* = \left(\sum_{j=1}^{n} (\, v_{ij} - v_j^*)^2 \right)^{\frac{1}{2}}$$
$$S_i^- = (\sum_{j=1}^{n} (\, v_{ij} - v_j^-)^2)^{\frac{1}{2}} \quad , \quad i = 1, \dots, m \qquad (5)$$

- Fifth step: Calculate the relative closeness to the ideal solution ($0 \le C_i^* \le 1$).

$$C_i^* = S_i^- / (\, S_i^* + S_i^-) \quad , \quad i = 1, \dots, m \qquad (6)$$

- Sixth step: Rank the alternatives in descending order of C_i^* or select alternatives with maximum value of C_i^*.

IV. Sequences and Distance Measures

Sequences are an important type of data which occur frequently in many domains, e.g. scientific, medical, security, and business applications. In addition, they can be used in controlling and tracking the history of daily activities. In many cases, sequences should be compared, thus distance measures would be needed. Sequence distance functions can be used for several applications e.g. clustering. They can also be applied in finding how a new sequence is similar to a known sequence. Several distance functions, such as character based, feature based, and conditional probability distribution based, have been proposed [14]. Edit distance is an example of character based distance measure, and d2 is a feature based one [12]. While these two measures are not proper choices in measuring the similarity of sequences, conditional probability distribution based distance gives acceptable results [14].

The conditional probability distribution (CPD) based distance uses the CPD of the next symbol, i.e. the symbol right after a segment of some fixed length L [14]. The resulting value can be used to characterize the structural properties of a given sequence. The difference between the corresponding CPDs is used to evaluate the similarity (or difference) between the two sequences. There are several methods to estimate this difference, but their time complexity is exponential in the length of the segment σ. There are alternative methods to avoid the computational burdens. According to Yang & Wang [14], given a sequence set S and the conditional probability distribution P modeling it, a sequence should subsume to a similar conditional probability distribution if the sequence can be predicted under P with relatively high probability. In the similarity measure defined by (7), P(s_i) is the probability of occurring the symbol s_i at any given position of any sequence in the database, and $P_S(\sigma)$ can serve as a measure of the similarity between the sequence σ and the sequence S. If $P_S(\sigma)$ is higher than the probability of predicting/generating σ by a memoryless random process, then we may infer that the sequence σ subsumes a similar CPD to that of S and may be considered as a member of S.

$$sim_s(\sigma) = \frac{P_S(\sigma)}{P^r(\sigma)} = \frac{\prod_{i=1}^{l} P_S(s_i \mid s_1 \dots s_{i-1})}{\prod_{i=1}^{l} P(s_i)} = \prod_{i}^{l} = 1 \frac{P_S(s_i \mid s_1 \dots s_{i-1})}{P(s_i)} \qquad (7)$$

The above measure is used only when the segment $s_1\, s_2 \dots s_{i-1}$ is a significant segment, i.e., $s_1\, s_2 \dots s_{i-1}$ occurs at least *c* times in a set of sequences. *c* is referred to as the significance threshold. Otherwise the longest significant suffix $s_j \dots s_{i-1}$ will be used in the estimation and the value of $P_S(s_i \mid s_j \dots s_{i-1})$ is supposed to be an estimation of $P_S(s_i \mid s_1 \dots s_{i-1})$. The value of CPD is also supposed to be uniformly distributed over the entire sequence. However, in some cases, especially when the sequence is long, there might be segments in the sequence having different values of CPDs. The problem can be resolved by modifying the above similarity measure to capture the maximum similarity between any segment of σ and S.

$$sim_s(\sigma) = \max_{1 \le j \le i \le l} sim_s(s_j \dots s_i) \qquad (8)$$

In CLUSEQ Algorithm [14], Probability Suffix Tree (PST) is used as an effective method for measuring this similarity. The algorithm starts with a null tree, and reverses the sequence. Then all suffixes of the reversed sequence are added to the tree as its nodes. A number and a probability distribution vector is assigned to each node, in which the number shows the frequency of the occurrence of its label and the probability distribution vector is used to preserve the CPD of the next symbol. A node with frequency greater than or equal to *c* is considered as significant node.

Probability estimation using this tree can be performed in two steps as follow:

- First step: Travers from the root along the path $\rightarrow s_{i-1} \rightarrow \ldots \rightarrow s_2 \rightarrow s_1$ and find a node labeled $s_j \ldots s_{i-1}$ as the longest significant suffix of $s_1 \ldots s_{i-1}$.
- Second step: Obtain probability of the symbol s_i.

For example, in Fig. 1 the thin line separates the set of significant nodes from the rest for c=2, and so does the bold line for c=4. The value of p(a|abb) can be estimated by starting at the root of the tree in Fig. 1, then the path →b→b→a is traversed until an insignificant node is observed, which happens at node z for c = 2. Therefore, the longest significant suffix of 'abb' is 'b' and the value of probability of 'a' which is stored in z will be retrieved. Hence p(a|abb) ≈p(a|b) =½. According to above explanations, the following equations (via a single scan of σ) is used to measure the similarity.

$$X_i = \frac{P_S(s_i \mid s_1 \ldots s_{i-1})}{P_S(s_i)} \qquad (9)$$

$$= \frac{C_S(s_1 \ldots s_{i-1}\, s_i)/C_S(s_1 \ldots s_{i-1})}{P_S(s_i)}$$

$$Y_i = \max\{\, Y_{i-1} * X_i \,, X_i\} \qquad (10)$$

$$Z_i = \max\{\, Z_{i-1} \,,\, Y_i\} \qquad (11)$$

Where Y_1=Z_1=X_1, and $C_S(\sigma)$ is the count of the segment σ.

V. Combining TOPSIS and Similarity Measure

Measuring similarity between sequences and ranking them is a MADM problem. Sequences and their elements are considered as alternatives and criteria, respectively. In this problem, especially when sequences are long there is a need for a more effective method. Here, TOPSIS as a well known MADM technique is used. We explain the method by the following example.

Let (A_1) = < (*aabb*) (*abbc*) (*cbbcabab*)> and (A_2) = < (*aabb*) (*abbacb*) (*cba*)> be the two sequences which should be compared. The criteria in sequence A_1 can be '*aabb'*, '*abbc'* and '*cbbcabab'*. The second elements of A_1 and A_2 are compared, for c=2. Firstly, probability suffix tree is built for '*abbc'* from A_1, then table 1 is completed by using (9), (10), (11). Thus, the similarity of '*abbc'* and '*abbacb'* is measured as 3.375. Similarly, the table is constructed for *'abbc'* and the result will be 2.25.

This process is repeated for the first and the third elements by constructing the tree for '*aabb*' and '*cbbcabab*', and then the table is completed for them. In the example, it is assumed that the probability of observing symbols a, b and c in the entire sequence database is 1/3.

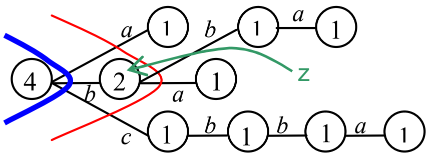

Fig 1. A Probabilistic Suffix Tree of 'abbc'

Table 1. Similarity Estimation of 'abbacb'

	a	***b***	***b***	***a***	***c***	***b***
$P_s(s_i \mid s_1 \ldots s_{i-1})$	1/4	1/2	1/2	1/2	0	1/2
X_i	3/4	3/2	3/2	3/2	0	3/2
Y_i	3/4	3/2	2.25	3.375	0	3/2
Z_i	3/4	3/2	2.25	3.375	3.375	3.375

We can obtain the similarity between the two sequences and construct decision matrix as follows.

c = 2	B_1	B_2	B_3
A_1	1	2.25	1.5
A_2	1	3.375	2.25

c = 4	B_1	B_2	B_3
A_1	1	2.25	3.375
A_2	1	2.25	2.25

In the next step we will rank the sequences by TOPSIS as follow. The method is implemented using two thresholds, i.e. c=2 and c=4, and the criteria are weighted equally, i.e., 1/3. Having decision matrices and knowing that the fact that the similarity between one element with itself must be greater than or equal to the similarity between that element and other ones, it is observed that the result in case c=4 is more acceptable. This result also shows the important role of c.

First step: Construct normalized decision matrix,

c = 2	B_1	B_2	B_3
A_1	$1/\sqrt{2}$	$2.25/\sqrt{16.4531}$	$1.5/\sqrt{7.3125}$
A_2	$1/\sqrt{2}$	$3.375/\sqrt{16.4531}$	$2.25/\sqrt{7.3125}$

c = 4	B_1	B_2	B_3
A_1	$1/\sqrt{2}$	$2.25/\sqrt{10.125}$	$3.375/\sqrt{16.4531}$
A_2	$1/\sqrt{2}$	$2.25/\sqrt{10.125}$	$2.25/\sqrt{16.4531}$

Second step: Obtain weighted normalized decision matrix,

c = 2	B_1	B_2	B_3
A_1	0.2357	0.1849	0.1849
A_2	0.2357	0.2773	0.2773

c = 4	B_1	B_2	B_3
A_1	0.2357	0.2357	0.2773
A_2	0.2357	0.2357	0.1849

Third step: Obtain positive ideal and negative ideal (the value of criteria shows the similarity and greater values are more desirable)

c = 2	c = 4
A^* = (0.2357 , 0.2773 , 0.2773)	A^* = (0.2357 , 0.2357 , 0.2773)
A^- = (0.2357 , 0.1849 , 0.1849)	A^- = (0.2357 , 0.2357 , 0.1849)

Forth step: Obtain distance of ith alternative from the ideals,

c = 2	c = 4
$S_1^* = \sqrt{0.017094005} = 0.130744$	$S_1^* = 0$
$S_1^- = 0$	$S_1^- =$ 0.0924
$S_2^* = 0$	$S_2^* =$ 0.0924
$S_2^- =$ 0.130744	$S_2^- =$ 0

Fifth step: Compute the relative closeness of *i*th alternative to the positive ideal solution,

$c = 2$	$c = 4$
$C_1^* = 0 \quad C_2^* = 1,$	$C_1^* = 1 \quad C_2^* = 0$

Sixth step: Rank the alternatives in descending order of C_i^*.

$c = 2$	$c = 4$
$A_2 > A_1$	$A_1 > A_2$

According to these computations, it is clear that acceptable result is achieved when *c*=4.

VI. Conclusion

This Paper considers the problem of measuring the similarity of sequences and ranking them, and proposes a new method. In the proposed method, sequence distance functions based on structural features are used. To obtain the structural features, probability suffix tree is used. The proposed method considers the ranking of sequences as a MADM problem and uses TOPSIS method in order to solve the problem. In this domain more general conditions can be assumed for sequences that can be taken into account in future studies.

References

[1] T. Chen, *Decision Analysis*, Publishing House for Science and Technology, Beijing, 1987.

[2] C. T. Chen, “Extensions of the TOPSIS for group decision-making under fuzzy environment”, *Fuzzy Sets and Systems*, 114 (2000), pp. 1–9.

[3] M.C. Chen, A.L. Chiu, H.H. Chang, “Mining changes in customer behavior in retail marketing”, *Expert Systems with Applications*, 28 (4) (2005), pp. 773–781.

[4] G. Fendel, J. Spronk, *Multiple Criteria Decision Methods and Applications*, Spring-Verlag, New York, 1983.

[5] C. L. Hwang, K. Yoon, *Multiple attributes decision making methods and applications*, Springer: Berlin Heidelberg, 1981.

[6] Y. J. Lai, C. L. Hwang, *Fuzzy multiple objective decision making methods and applications*, New York: Springer-Verlag, 1994.

[7] D. Li, “Multiattribute decision making models and methods using intuitionistic fuzzy sets”, *J. Comput. System Sci.*, 70 (2005), pp. 73–85.

[8] D.F. Li, J.B. Yang, “Fuzzy linear programming technique for multi-attribute group decision making in fuzzy environments”, *Information Sciences*, vol.158 (2004), pp. 263–275.

[9] P. Nijkamp, “A Multi-criteria Analysis for Project Evaluation: Economic-Ecological Evaluation of a Land Reclamation Project”, *Papers of the Regional Science Association*, Vol. 35, No. 1, (1974), pp. 87–111.

[10] H.S. Shih, H.J Shyur, E.S. Lee, “An extension of TOPSIS for group decision making” *Mathematical and Computer Modelling,* 45 (2007), pp. 801–813.

[11] C-Y. Tsai, Y - C Shieh, “A change detection method for sequential patterns”, *Decision Support Systems*, 46 (2009), pp. 501–511.

[12] D. C. Torney, C. Burks, D. Davison, K. M. Sirotkin, “Computation of d2: A Measure of Sequence Dissimilarity”, *Computers and DNA*, 1990, pp. 109–125.

[13] K. P. Yoon, Wang, C.L. Hwang, *Multiple Attribute Decision Making*: *An Introduction*, Sage University Papers, 1995.

[14] J. Yang, W. Wang. “CLUSEQ: efficient and effective sequence clustering”. *Proceedings. 19th International Conference on Data Engineering*, 2003, pp. 101–112.

[15] Amir Yousefli, Majeed Heydari, Kamran Shahanaghi “Development of Linear Programming Technique for Multidimensional Analysis of Preference in Fuzzy Environment”, *Journal of Uncertain Systems*, Vol.3, No.2 (2009), 108–113.