Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230202

Extracting Temporal Relationships in EHR: Application to COVID-19 Patients

Carlos MOLINA[a,1] and Belén PRADOS-SUAREZ[a]
[a] *Department of Software Engineering, University of Granada, Spain*

Abstract. Association rules are one of the most used data mining techniques. The first proposals have considered relations over time in different ways, resulting in the so-called *Temporal Association Rules (TAR)*. Although there are some proposals to extract association rules in OLAP systems, to the best of our knowledge, there is no method proposed to extract temporal association rules over multidimensional models in these kinds of systems. In this paper we study the adaptation of TAR to multidimensional structures, identifying the dimension that establishes the number of transactions and how to find time relative correlations between the other dimensions. A new method called *COGtARE* is presented as an extension of a previous approach proposed to reduce the complexity of the resulting set of association rules. The method is tested in application to COVID-19 patients data..

Keywords. Temporal Association Rules (TAR), Multidimensional Model, Complexity, COVID-19

1. Introduction

Association rules are one of the most commonly used methods for decision making. Since the first proposal by Agrawal et al. [1], several extensions of the concept have been proposed. Tung et al. [2] extended the association rules approach considering time relationships between records – what they call *inter-transaction association rules*. The authors proposed an algorithm to extract relations as follows:

when A appears, then B appears T later,

where *A* and *B* are a set of items, and *T* is a measure of time (e.g. *3 days*, *1 month*, etc.). Lu et al. [3] extended the model from one-dimensional inter-transaction (time) to N-dimensional inter-transaction association rules (e.g. time and distance) with application to stock movement prediction. The time concept in association rules has been studied from several perspectives. An exhaustive review and classification can be found in [4].

Association rules have been adapted to other systems that differ from the transactional one. An example are OLAP (On-line Analytical Processing) systems where the data are organized using multidimensional structures called datacubes [5]. One common characteristic of these datacubes is that there is always a dimension to represent time, since OLAPs are used for strategic analysis in organizations. To the best of our knowledge, although there are proposals for association rules over datacubes (e.g. [6,7]), none include time in their methods.

[1] Corresponding Author: Carlos Molina, E-mail: carlosmo@ugr.es.

In this paper we propose a new method to extract temporal association rules from fuzzy datacubes adapted to these analysis-oriented systems. There is no standard for OLAP structures, so we first need to establish the multidimensional model we will use to represent the data. In our case, we use a Fuzzy Multidimensional Model (see [8] for details). As starting point, we will describe an association rules method according to the complexity defined for this structure [7] aimed at getting understandable association rule sets (Section 3.1). The next section presents the datacube applied to COVID-19 patients' data to test the proposed method. The last one presents the main conclusions.

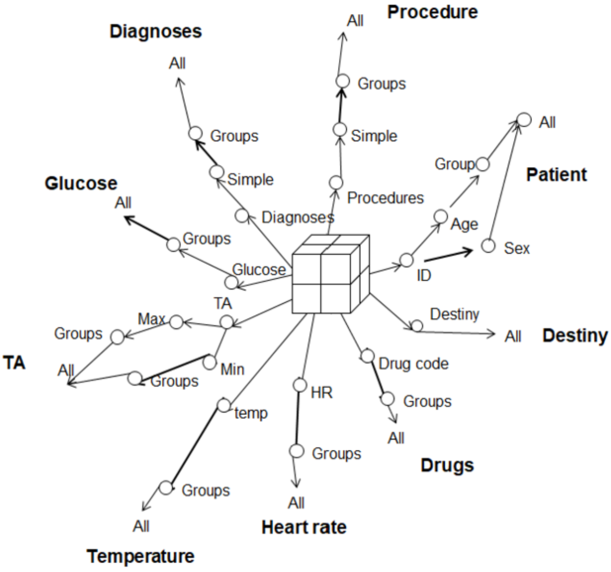

Figure 1. Multidimensional model for COVID-19 patients data (9).

2. Temporal Association Rules over Fuzzy Datacubes

We first need to present a datacube applied to COVID-19 patients' data to test and illustrate the method. Figure 1 shows the multidimensional model. In this example, we have patients' information on COVID-19. The *Patient* dimension includes data of the patients that do not change over time (e.g. age, sex, etc.). In our studies we want to know how the relationships between the other variables evolve over time for each patient, so this dimension will establish the number of records we have (for each patient we will have times series for the rest of the variables and we want to identify temporal relationships common to all the patients). We will refer to these kinds of dimensions whose values do not change over time as *fixed dimensions (D_F)*, which will allow us to identify each of the entities over time.

As we have mentioned before, in most datacubes there is dimension that represents time, which is used for evolution analysis to see trends in the rest of the variables or dimensions. In this example the time dimension is *Time*. This dimension establishes the period in which we have values for the rest of the variables.

The rest of the dimensions have values that change over time for each patient, and we want to find relationships related to time between them. We call these *variable dimensions (D_V)* (e.g. *Drugs* dimension). The last element we need is a way to measure the representativeness and strength of the relationship. For the former, we use the normal measure in association rules: the *Support* (in this example the number of patients that satisfy the rule) and the *Certainty Factor (CF)* [10] instead of *Confidence* because of the known problems with very frequent items [11].

Now we have all the elements to define the *Temporal Association Rules* over datacubes.

Definition. The temporal association rules AR_T are

$$AR_T=(a\cdot b \rightarrow b',t,Support=\alpha,CF=\delta)$$

where:

- $\forall i \in a/i \in D_F$, *i* are items belonging to the *Fixed Dimensions.*
- $\forall i \in b/i \in D_V$ and $\forall i \in b'/i \in D_V$, *$* and *i'* are items from the *Variable Dimensions*.
- *t* is a time measure.
- *Support=α* is the representativeness of the rule, where $\alpha \in [0,1]$.
- *CF=δ* is the strength of the relationship, where $\delta \in [-1,1]$.

which means that for a concrete value of time dimension I_T where $a\cdot b$ appears, $a\cdot b'$ appears in the moment I_T+t.

3. Methods

Once we have formally defined the temporal association rules, we present the algorithm to extract them from datacubes. We adapted a previous method (COGARE) that extracts association rules reducing the complexity of the results. The method uses the hierarchies of the dimensions in the datacubes to reduce the number of rules, presenting concepts that are more understandable by users. The method extracts rules over datacubes and diminishes complexity by using the concepts defined in the hierarchy over each dimension to reduce the number of rules and improve interpretability. It comprises three main stages:

- *Itemset generation*: the algorithm uses the Apriori algorithm [1] adapted to the multidimensional model. It takes into account the hierarchy of the dimension when an itemset is not frequent and looks for a generalization that may be frequent (bottom-up approach). In this generalization the support threshold is adapted such that the more general the items are, the higher the threshold is (see [7] for details).
- *Rules generation*: algorithm extracts rules using an Apriori-like approach [1].
- *Rules generalization*: the rule set obtained in the previous stage is generalized to reduce the complexity of the result. In this step, the algorithm tries to generalize the elements in the association rules by defining the items at a higher level (more abstract). If the generalized rules include others (defined over more concrete values but representing the same knowledge) those are deleted. On each step, the quality of the rule set is controlled so if it decreases down to an established threshold, the operation is not applied.

As mentioned before, the associations rules obtained with this process are pruned according to a *certainty factor (CF)* instead of *confidence* to avoid some of the well-known problems of this quality measure.

3.1. COGtARE: COmplexity Guided Temporal Association Rule Extraction

In this section we present the changes made to the COGARE method to extract temporal association rules. As we have mentioned, the concept of support is different. Now we have to calculate the number of different values (records) in the *fixed dimensions*. This can be done querying the datacube applying slice (selecting only the *fixed dimensions*) and applying the values in *variable dimensions* as restrictions.

Once we have the frequent itemsets, next step is the rule generation process. In our approach we have combined two itemsets, one for antecedent and the other for consequent items. But not all itemsets can be combined – they have to be compatible in the sense that both represent the same entity. In our case, the same entity means the itemsets share the same values for the *fixed dimensions*:

Definition. In I=$a{\cdot}b$ and I'=$a'{\cdot}b'$ we have two itemsets where 1) $\forall i \in a/i \in D_F$, and $\forall i' \in a'/i' \in D_F$, 2) $\forall i \in b/i \in D_V$, and $\forall i' \in b'/i' \in D_V$, then I and I' are compatible if a=a'.

The last aspect is how we calculate the time relation (*t*) part of the rules. To represent it, we consider the period between the two closest records that satisfy the restrictions (b and b' in the rule) for all the entities that meet the rule. With all these values, the user has a 95% confidence interval. The rest of the steps do not change, so we have all the elements to extract temporal association rules over the datacubes. In next section we test the method over the running examples datacube over COVID-19 patients' data and shows some interesting extracted rules. With this definition, we have executed the proposed algorithm over the datacube in Figure 1. Table 1 shows some examples of extracted rules.

Table 1. Examples of rules obtained.

Rules	**Support**	**CF**	**t in days (avg and interval)**
{}·{Temperature is Normal, Drug N02BE01, Drug V03AN01} → {Destiny is Home}	0.11	0.25	4.3 [3.4,5.2]
{}·{Procedure 0BJ0XZZ applied, Drug D08AC02, Drug V03AN01}→ {Destiny is Home}	0.11	0.33	6.8 [5.6,8.1]
{}·{TA max is high, Drug A02BC01} → {Drug D08AC02}	0.68	0.32	1.3 [1.1,1.5]
{}·{Procedure 0BJ0XZZ applied, Drug V03AN01}→{Destiny is Home}	0.12	0.30	7.2 [5.8,8.5]

4. Conclusions

In this paper we have introduced the time relationships in the extraction of association rules over multidimensional structures in OLAP systems. We have the different roles that dimensions in datacube may play, where we have the ones that identify the entities and which do not change over time (*fixed dimensions*), the dimensions that do change over time, which is where we can find the temporal correlations (*variable dimensions*), and the dimensions that measure the passage of time (*Time dimension*). We have proposed a method to deal with these categories and properly calculate the support of the item sets and the quality of temporal association rules.

In this paper we have applied the proposed method over COVID-19 patient’s data.

Acknowledgement

This research is partially supported by B-TIC-744-UGR20 *ADIM: Accesibilidad de Datos para Investigación Médica* of the Junta de Andalucía.

References

[1] Agrawal R, Sritkant R. Fast Algorithms for Mining Association Rules in Large Databases. In: Proceedings of 20th International Conference on Very large data Bases. 1994. p. 478–99.

[2] Tung AKH, Lu H, Han J, Feng L. Breaking the barrier of transactions. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '99. ACM Press; 1999.

[3] Lu H, Han J, Feng L. Stock movement prediction and n-dimensional inter-transaction association rules. In: 1998 ACM SIGMOD Workshop on Research Issues on DMi and KD, Seattle, WA, USA, ACM, New York, USA. 1998.

[4] Segura-Delgado A, Gacto MJ, Alcalá R, Alcalá-Fdez J. Temporal association rule mining: An overview considering the time variable as an integral or implied component. WIREs Data Min Knowl Discov. 2020 Apr;10(4).

[5] Codd EF. Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate. E.F. Codd and Associates; 1993.

[6] Zhu H. On-Line Analytical Mining of Association Rules [Internet] [PhD Thesis]. Simon Fraser University; 1998.

[7] Marín N, Molina C, Serrano JM, Vila A. A Complexity guided algorithm for association rule extraction on fuzzy datacubes. IEEE Tras Fuzzy Syst. 2008;16:693–714.

[8] Molina C, Rodriguez-Ariza L, Sanchez D, Vila MA. A New Fuzzy Multidimensional Model. IEEE Trans Fuzzy Syst. 2006 Dec;14(6):897–912.

[9] MH Hospitales. Covid Data Save Lives [Internet]. 2021. Available from: https://www.hmhospitales.com/coronavirus/covid-data-save-lives/english-version

[10] Shortliffe E, Buchanan B. A model of inexact reasoning in medicine. Math Biosci. 1975;23:351–79.

[11] Delgado M, Marin N, Sanchez D, Vila MA. Fuzzy Association Rules: General Model and Applications. IEEE Trans Fuzzy Syst. 2003;11(2):214–25.