The Sequential Associative Rules Analysis of Patient's Physical Characteristics

Nataliya Shakhovska [1][0000-0002-6875-8534], Roman Holoshchuk [1] [0000-0002-1811-3025],

Solomia Fedushko [1] [0000-0001-7548-5856], Oleh Kosar [1],

Roman Danel [2][0000-0001-9333-0567], Michal Řepka [2]

[1]Lviv Polytechnic National University, Lviv, Ukraine

[2]Institute of Technology and Businesses in České Budějovice, Czech Republic

nataliya.b.shakhovska@lpnu.ua, roman.o.holoshchuk@lpnu.ua, rdanel@mail.vstecb.cz, repka@mail.vstecb.cz

Abstract. The sequential associative rules method creating are described. The difference between classical associative rules and sequential associative rules is given. The patient medical data is analyzed. The main associative rules characteristics are given. For modified AprioriTID method unique identifier for the patient set of patient analyzes has been entered. Additional numerical attributes of the investigated objects are indicated. The distinction between associative rules and sequential analysis is given. The analysis of the results of well-known methods and developed method is given.

Keywords: sequential associative rules, artificial intelligence, support, confidence, sequential analysis

1 Introduction

Association rules are a data mining technique used to discover frequent patterns in a data set. In this work, association rules are used in the medical domain, where data sets are generally high dimensional. The chief disadvantage about mining association rules in a high dimensional data set is the huge number of patterns that are discovered, most of which are irrelevant or redundant. This disadvantage is grown when Big data is used. The multidimensional view of the data is well used for data visualization and analysis tasks, but due to the hypercube dissipation, the amount of data in this case is greater than the relational representation that is not acceptable to the Big Data. Object representation allows you to store object in the form of attributes, their characteristics and relationships between characteristics. For some modification, it can be used for Big Data.

In medical and biological research, as well as in practical medicine, the range of tasks to be solved is so wide that it is possible to use any of the methodologies of Data

Mining. An example can be the construction of a diagnostic system or the study of the effectiveness of surgical intervention [1 – 3].

One of the most advanced areas of medicine is bioinformatics. The object of bioinformatics research is huge amounts of information about DNA sequences and the primary structure of proteins that arose as a result of studying the structure of genomes of microorganisms, mammals and humans. Abstracted from the specific content of this information, it can be regarded as a set of genetic texts, consisting of extended character sequences. Detection of structural laws in such sequences is a number of tasks, effectively solved by means of Data Mining, for example, by means of sequencing and associative analysis [4 – 5].

The purpose of the study is to identify the most important rules for constructing associative rules. We should analyze not only single parameters and theirs values but also combining of these parameters in groups. Determination of the patterns of constructing associative rules and the division of physical indicators at different levels of the hierarchy.

2 Objects and methods of research

One of the most common data analysis tasks is to identify sets of objects that are often encountered in a large set of objects [5 – 8]. We describe this problem in a generalized form. To do this, we denote the objects that make up the study sets (itemsets), as follows [9 – 10]:

$$I = \{i_1, i_2, \ldots, i_j, \ldots, i_n\}, \quad (1)$$

where i_j – objects included in the studied sets; n – total number of objects.
In the field of medicine, such objects, for example, are indicators and analyzes of the patient (Table 1).

Table 1. Objects included in the study set

Id	Indicator	Value
0	Blood pressure	120/80 mm. Hg.
1	Venous pressure	70 mm. H_2O
2	Capillary pressure	70 mm. Hg.
3	Pulse	85 beats/min
4	Temperature	36,6 C
5	Level of hemoglobin in the blood	145 Hb
6	pH	7.35

In this way they correspond to the following set of objects:

```
I = {arterial pressure, venous pressure, capillary pres-
sure, pulse, temperature, hemoglobin level in blood, pH}.
```

Sets of objects from the I set, stored in a database and subject to analysis, are called transactions. We describe the transaction as a subset of the set I:

$$T = \{i_j | i_i \in I\}. \quad (2)$$

Such transactions in the hospital are in accordance with the delivery of medical examinations of the patient and stored in the database in the form of a medical card. They list the tests that the patient passed for a history and diagnosis.

The set of transactions, the information about which is available for analysis, will be described by the following set:

$$D = \{T_1, T_2, \dots, T_r, \dots, T_m\}, \quad (3)$$

where *m* – the number of transactions available for analysis.

3 Research results

To use Data Mining methods, the set D can be represented as a table (Table 2).

The set of transactions, which includes j_i objects, is indicated as follows [7]:

$$D = \{T_r | i_j \in T_r;\ j = 1..n;\ r = 1..m\} \subseteq D \quad (4)$$

In this example, the set of transactions containing the Object Temperature is the following:

Table 2. A set of investigated objects

Transaction number	Indicator number	Indicator	Value
0	0	Blood pressure	110/75 mm. Hg.
0	3	Pulse	110 beats/min
0	1	Venous pressure	58 mm. H_2O
1	4	Temperature	37.4 °C
1	5	pH	7.46
2	1	Venous pressure	72 mm. H_2O
2	6	pH	7.81
2	4	Temperature	37.2 °C

In this example, the set of transactions containing the Object Temperature is the following set:

```
Dtemperature = {{Temperature, pH },
                {Venous pressure, pH, Temperature}}
```

Some arbitrary set of objects (itemset) is denoted as follows:

$$F = \{i_j | i_j \in I;\ j = 1..n\}. \quad (5)$$

The set of transactions that includes the set F is denoted as follows:

$$D_F = \{T_r \mid F \subseteq T_r;\ r = 1..m\} \subseteq D. \quad (6)$$

The ratio of the number of transactions, which includes the set F, to the total number of transactions is called support of the set F and denoted by Supp (F):

$$Supp(F) = |D_F|/D. \quad (7)$$

For example, for a set `{pH, temperature}` the subtraction will be equal to 2/3, because this set is included in two transactions (numbers 1 and 2) of the three possible.

When searching, an analyst can specify the minimum value of maintaining interesting sets – Suppmin. A set is called large if its value exceeds the minimum support value specified by the user:

$$Supp(F) > Supp_{min}. \quad (8)$$

So, when searching for associative rules you need to find the set of all frequent sets:

$$L = \{F | Supp(F) > Supp_{min}\} \quad (9)$$

In this case, the sets with Suppmin = 2/3 are the following:

```
{Venous pressure} Suppmin = 2/3;
{Temperature} Suppmin = 2/3;
{pH, Temperature} Suppmin = 2/3;
```

In an analysis, the sequence of events is often of interest. When detecting regularities in such sequences, it is possible to predict with some degree the occurrence of events in the future, which allows us to make more correct decisions. A sequence is called an ordered set of objects. To do this, the order must be given to the set [8].

Then the sequence of objects can be described as follows:

$$S = \{\ldots, i_p, \ldots, i_q\}, \text{where } p < q. \quad (10)$$

For example, in the case of analyzes such a sequence of objects may be the date of delivery of analyzes. Such a sequence:

```
S = {(hemoglobin level, 10.10.2017),
     (venous pressure, 09/25/2017),
     (pH, 28.09.2017)}
```

can be interpreted as a sequence of delivery of tests by one person at different times (initially measured venous pressure, then measured the pH level, and finally the level of hemoglobin).

There are two types of sequences: with cycles and without cycles. In the first case it is allowed to enter the sequence of the same object at different positions:

$$S = \{\dots, i_p, \dots, i_q, \dots\}, \text{where } p < q, i_q = i_p . \quad (11)$$

It is said that transaction T contains the sequence S, if $S \subseteq T$ and the objects included in S, also belong to the set of T, with preservation of the relation of order. It is supposed that in the set T between objects in the sequence of S there may be other objects.

The maintenance of the sequence S is the ratio of the number of transactions, which includes the sequence of S, to the total number of transactions. The sequence is frequent if its support exceeds the minimum support given by the user:

$$\text{Supp}(S) > \text{Supp}_{\min} . \quad (12)$$

The task of sequential analysis is to search all frequent sequences:

$$L = \{S|\text{Supp}(S) > \text{Supp}_{\min}\} . \quad (13)$$

The main difference between the problems of sequential analysis from the search for associative rules is to establish a relation of order between objects of the set I. This relation can be determined in different ways. In the analysis of the sequence of events occurring in time, the objects of the set I are events, and the order of relationships corresponds to the chronology of their appearance. For example, analyzing sequences of assays in a hospital are sets of analyzes that the patient submits at different times, and the order of reference is the time of the implementation of these analyzes.

```
D = {{(temperature, blood pressure, capillary pressure),
      (pH, temperature, pulse)},
     {(hemoglobin level in blood, temperature),
      (blood pressure, temperature),
      (temperature, venous pressure)},
     {( hemoglobin level in the blood)}}.
```

Of course, there is a problem of identification of patients. In practice, this is decided by the introduction of medical cards that have a unique identifier (table 3).

Table 3. A unique identifier for the set of analyzes

Patient ID	Sequence of analyzes delivery
0	(temperature, arterial pressure, capillary pressure), (pH, temperature, pulse)
1	(hemoglobin level in the blood, temperature), (blood pressure, temperature), (temperature, venous pressure)
2	(hemoglobin level in the blood)

The following sequence can be interpreted as follows: the patient with the ID 0 initially passed the temperature, the arterial and capillary pressure, and then passed the pH, temperature and pulse rate with his visit. For example, the support for the `{(blood pressure, temperature)}` sequence is 2/3, since it is found in patients with identifiers 0 and 1.

In many applications, objects of the set I naturally combine into groups that in turn can also be grouped into more general groups, etc. Thus, the hierarchical structure of objects is obtained.

An example of such a hierarchy may be the following categorization of analyzes:

```
Pressure:
· Arterial;
· Venous;
· Capillary
Physical indicators:
· Temperature
Blood test:
· Hemoglobin level;
· PH
```

The presence of a hierarchy changes the perception of when an object i is present in transaction T. Obviously, support is not a separate object, but the group to which it is included is greater:

$$\mathrm{Supp}(\mathrm{I_q}) \geq \mathrm{Supp}(\mathrm{i_j}) \tag{14}$$

where $i_j \in Iq$.

This is due to the fact that when analyzing groups, not only transactions that include a separate object, but also transactions containing all objects of the analyzed group are counted. For example, if `Supp {blood pressure, temperature} = 2/3`, then support `Supp {pressure, physical parameters} = 2/3`, since the objects of the groups of pressure and physical parameters are included in the transaction with the identifiers 0 and 1.

Using the hierarchy allows you to determine the connection that goes into higher levels of the hierarchy, since the support for the set can increase if the entry of the group, and not its object, is counted. In addition to the search for kits that often occur in transactions, which in turn consist of objects $F = \{i | i \hat{I} I\}$ or groups of the same level of the hierarchy:

$$\mathrm{F} = \{\mathrm{I^g} | \mathrm{I^g} \in \mathrm{I^{g+1}}\}. \tag{15}$$

You can also consider mixed sets of objects and groups:

$$\mathrm{F} = \{\mathrm{i}, \mathrm{I^g} | \mathrm{i} \in \mathrm{I^g} \in \mathrm{I^{g+1}}\}. \tag{16}$$

This allows you to extend the analysis and gain additional knowledge.

In the hierarchical structure of objects, you can change the nature of the search by changing the analyzed level. Obviously, the more objects in the set I, the more objects in transactions T and frequent sets. This in turn increases search time and complicates the analysis of results. You can reduce or increase the amount of data using the hierarchical representation of the objects under analysis. Moving up the hierarchy, we summarize the data and reduce their number, and vice versa.

The disadvantage of generalizing objects is the less usefulness of the knowledge gained, since in this case they relate to groups that do not always have useful information. To achieve a compromise between group analysis and analysis of individual objects, they often do the following: first analyze the groups, and then, depending on the results, investigate the objects that interest the group analyst. In any case, it can be argued that the presence of a hierarchy in objects and its use in the task of finding associative rules allows you to perform a more flexible analysis and gain additional knowledge.

In the considered problem of searching for associative rules, the presence of an object in a transaction was determined only by its presence in it ($i_j \in T$) or the absence ($i_j \notin T$). Often, objects have additional attributes, usually numeric. For example, analyzes in a transaction have attributes: value and duration. In this case, the presence of an object in the set can be determined not only by the fact of its presence, but also the execution of the condition in relation to a certain attribute. For example, in analyzing transactions performed by patients, they are interested in not only the value of the analysis, but also in how well this indicator is stable (long-term).

You can add additional objects to explore the sets in order to extend the analysis capabilities by searching for associative rules. In the general case, they may have a nature different from the main objects. For example, in the case of delivery of tests, you can enter the field of delivery frequency or symptoms that precede the delivery of these particular analyzes.

Solving the problem of finding associative rules, as well as any task, is to process the output and obtain the results. A certain Data Mining algorithm performs processing of the initial data.

The results obtained in solving this problem are accepted in the form of associative rules. In this regard, when searching for them, there are two main stages:

1. Finding all large sets of objects;
2. Generation of associative rules from found large sets of objects.

Associative rules are as follows:

```
If (condition) then (result),
```

where condition is usually not a logical expression (as in the classification rules), but a set of objects from the set I, with which associated (associated) objects are included in the result of this rule.
For example, associative rule:

```
If (blood pressure, pH) then (hemoglobin level)
```

means that if the patient is measured by arterial pressure and pH level, he also measured by hemoglobin level.
As already noted, in associative rules the condition and the result are objects of the set I:

```
If X then Y,
```

where $X \in I, Y \in I, X \cup Y = \varphi$.

The main advantage of associative rules is their easy perception by a person and a simple interpretation of programming languages. However, they are not always useful. There are three types of rules:

1. Useful rules – contain valid information that was previously unknown but has a logical explanation. Such rules can be used for making decisions that are beneficial;
2. Trivial rules – contain valid and easily understandable information that is already known. Such rules, although they can be explained, but cannot bring any benefits, as they reflect or known laws in the studied area, or the results of past activity. Sometimes such rules can be used to verify the implementation of decisions taken on the basis of preliminary analysis;
3. Unclear rules – contain information that cannot be explained. Such rules can be obtained either based on abnormal values, or deeply hidden knowledge. Directly such rules cannot be used for decision making, since their lack of clarity can lead to unpredictable results. For better understanding, further analysis is required.

Associative rules are built on the basis of large sets. So, the rules built on the basis of the set F, are all possible combinations of objects included in it.

For example, for the set `{arterial pressure, temperature, pulse}` the following associative rules can be constructed:

```
If (arterial pressure) then (temperature);
If (arterial pressure) then (pulse);
If (arterial pressure) then (temperature);
If (arterial pressure) then (temperature, pulse);
If (temperature, pulse) then (arterial pressure);
```

And so on.

Thus, the number of associative rules can be very large and bad for human perception. In addition, not all of the built-in rules carry useful information. To assess their usefulness, the following values are entered:

- Support – shows which percentage of transactions supports this rule (we found rules, where Support is upper then 75%).
- Confidence – shows the probability that the presence of a set Y in the transaction in the set X implies (we found rules, where Confidence is upper then 0.5).
- Improvement – indicates whether this rule is useful for research.

These estimates are used when generating rules. An analyst when searching for associative rules specifies the minimum values of these variables. As a result, those rules that do not satisfy these conditions are discarded and are not included in the solution of the problem.

If objects have additional attributes that affect the composition of objects in transactions, and therefore in sets, then they should be taken into account in generated rules. In this case, the conditional part of the rules will not only include verification of the existence of an object in a transaction, but also more complex comparing operations:

more, less, includes, etc. The resulting part of the rules may also contain statements about the attribute values. For example, if an indicator is considered topical, then the rules may look like this:

```
If pH.relevance > 10 days then the level of hemoglobin in
the blood.relevance  < 3 days.
```

This rule states that the patient did the pH analysis more than 10 days ago, then probably his analysis of hemoglobin in the blood is valid for no more than 3 days.

The rules are stored to XML documents for further processing [8]. The XML documents could be static or dynamic. The main differences between static and dynamic XML documents are:

- Availability of validity period

A static XML document does not contain elements that indicate the expiration date of this document. In contrast, a dynamic XML document initially contains at least one element that indicates the validity period of a particular version of the document.

- Persistence of displayed information

Once created, the information of a static XML document remains valid at all times. Conversely, the version of the dynamic XML document is valid only for the period specified in the corresponding elements. As soon as a new version appears, the information contained in the previous version is replaced.

Most of the work on finding associative rules in static XML documents is related to the use of XML-based algorithms based on the Apriori algorithm. However, there are a number of other approaches.

Table 4. Representation of static xml document

```
<?xml version="1.0" encoding="UTF-8"?>
<patient>
  <name>Tom Johnson</name>
  <street>Bandery</street>
  <city>Lviv</city>
  <analyse>
  <analyse_date>15/01/2017</analyse_date>
    <results>
  <temperature>38.2</temperature>
  <venouse_pressure>72</venouse_pressure>
      <pulse>110</pulse>
      <pH>7.46</pH>
    <hemoglobin>145</hemoglobin>
    </results>
  </analyse>
</patient>
```

The number of useful dependencies found by different methods from the volume of the analyzed data (Table 5). The comparison is made between Apriori [10], FP-tree [11, 12] and proposed method. This methods implementation is done using RStudio.

Table 5. The number of useful dependencies found by different methods

Amount of records	Proposed method	FP-tree	Apriori
200	28	21	17
400	42	36	18
600	56	45	26
800	59	51	32

The comparison results show that proposed method is not dominated by well-known Apriori ans FT-tree.

The developed algorithm makes it possible to assert that the problem of identifying sequential associative rules belongs to the class of P-problems. So, the search algorithm for sequential associative rules is well resolved with MapReduce.

In addition to several successive implementations, parallel realities for working with Big data are not widely available. One example of a batch implementation is a well-known statistical computation with the package R, called "arules".

Parallel implementation of the FP-Groth program is available in the library for studying the open-source computer (MLlib) Apache Spark and Apache Mahout.

Conclusion

The task of finding associative rules is to identify sets of objects that are commonly encountered in a large number of objects. The task of sequential analysis is to search for frequent sequences. The main difference between the tasks of sequential analysis from the search for associative rules is to establish a relationship of order between objects.

The presence of a hierarchy in objects and its use in the task of finding associative rules allows you to perform a more flexible analysis and obtain additional knowledge. The results of the solution of the problem are presented in the form of associative rules [13, 14], conditional and the final part of which contains sets of objects.

References

1. Perova, I., Bodyanskiy, Y.: Fast Medical Diagnostics Using Autoassociative Neuro-Fuzzy Memory. International Journal of Computing, 16(1), 34-40 (2017)
2. Tkachenko, R., Doroshenko, A., Izonin, I., Tsymbal, Y., & Havrysh, B.: Imbalance Data Classification via Neural-Like Structures of Geometric Transformations Model: Local and Global Approaches. In: International Conference on Theory and Applications of Fuzzy Systems and Soft Computing, pp. 112-122, Springer (2018).

3. Syerov, Y., Shakhovska, N., Fedushko S.: Method of the data adequacy determination of personal medical profiles (in press).
4. Korobiichuk, I., Fedushko, S., Juś, A., Syerov, Y.: Methods of Determining Information Support of Web Community User Personal Data Verification System. Automation 2017. Advances in Intelligent Systems and Computing, vol 550. Springer, pp-144-150 (2017).
5. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems, 30(1-7), 107-117 (1998).
6. Negnevitsky, M.: Artificial intelligence: a guide to intelligent systems. Pearson (2005).
7. Jain, V., Benyoucef, L., Deshmukh, S. G.: A new approach for evaluating agility in supply chains using fuzzy association rules mining. Engineering Applications of Artificial Intelligence, 21(3), 367-385 (2008).
8. Shakhovska, N., Kaminskyy, R., Zasoba, E., Tsiutsiura, M.: Association rules mining in big data. International Journal of Computing, 17(1), 25-32 (2018).
9. Porkodi, R., Shivakumar, B.L.: An improved association rule mining technique for xml data using Xquery and Apriori algorithm. Advance Computing, 1510-1514 (2009).
10. Woo, J.: Apriori-Map/Reduce algorithm. In: Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications, p. 1, (2012).
11. Hunyadi, D.: Performance comparison of Apriori and FP-Growth algorithms in generating association rules. European Computing Conference, 376-381 (2011).
12. Khurana, K., Sharma, S.: A comparative analysis of association rule mining algorithms, International Journal of Scientific and Research Publications, vol. 3, issue 5 (2013).
13. Boyko, N., Sviridova, T., Shakhovska, N.: Use of machine learning in the forecast of clinical consequences of cancer diseases. In: 7th Mediterranean Conference on Embedded Computing (MECO), pp. 1-6 (2018).
14. Veres, O., Shakhovska, N.: Elements of the formal model big date. In: XI International Conference on Perspective Technologies and Methods in MEMS Design (MEMSTECH), pp. 81-83 (2015).