Association Rule Mining Methods as Means of Forming the System of Life Quality Indicators

Luidmila P. Bilgaeva[1], Erzhena Ts. Sadykova[2], Gregory V. Badmaev[1]

[1] East Siberia State University of Technology and Managemen, Ulan-Ude, Russia,
http://www.esstu.ru
[2] Baikal Institute of Nature Management SB RAS, Ulan-Ude, Russia,
http://www.binm.ru

Abstract. The paper is devoted to comparing of the association rules mining methods and choosing the best method for the formation of the indicators system that affects the quality of life. Three methods are considered: AprioriTiD, FPG and ABBM. For each method, such metrics are calculated as support, confidence, lift and running time, the values of which allow you to discover the best method. It results in the extraction of useful association rules showing how life quality indicators are related to each other. Later on it can be used to solve the problems of analyzing and forecasting.

Keywords: data mining, association rule mining, support, confidence, lift, frequent itemsets, truncation of candidates, algorithm running time, life quality indicators

1 Introduction

Association rules mining is one of the modern technology tasks for Data Mining, which involves finding the patterns between some related events, the definition of interrelated objects and their location in the state space [13]. Initially, the methods of association rules mining were created to assess the consumer basket [2]. At present, the algorithms for mining of association rules are used to solve various problems associated with the identification of different regularities, for example, in medical diagnostics [5], pharmacology [11], studying of forest fires [8], etc.

In this paper, we propose to perform a comparative analysis of the association rules mining methods for the formation of a system of indicators characterizing the life quality of the population. To assess the life quality the number of indicators, such as socio-economic and environmental ones, are used. They allow you to determine the degree of satisfaction of personal material and social needs. In its turn, each indicator is characterized by a variety of factors affecting them. The association rules mining will allow forming a small number of indicators and factors, which has the greatest impact on the life quality definition. Further on, this approach can be used to solve problems of analyzing and forecasting of socio-economic processes.

As the initial data we use the classification of socio-economic and environmental indicators proposed in [12].

2 Valid method choice

To select the best method for association rules mining, the authors developed certain criteria and, in accordance with them, analyzed a certain number of methods. The results are given in Table 1.

Table 1. Comparative analysis of association rules mining methods

No	Methods	Criteria				
		Implementation simplicity	Small number of candidates	Small running time	ID transactions application	Candidates' truncation possibility
1	AIS	+	−	−	−	−
2	SETM	+	−	−	+	−
3	Apriori	+	+	+	−	+
4	AprioriSome	−	+	+	−	+
5	AprioriTID	+	+	+	+	+
6	FPG	−	+	+	−	+
7	ABBM	−	+	+	+	+

The table shows that the AIS and SETM methods generate a large number of candidates, but cannot truncate them. As a result, the running time of these algorithms increases significantly, which makes them less popular [7].

The Apriori, AprioriTID, AprioriSome, FPG and ABBM methods do not have this disadvantage because they generate a smaller number of candidates and are able to truncate candidates, which provides fewer loads on RAM.

The analysis shows that the Apriori and AprioriTid methods proposed in [2,3] are easy to implement, because to store intermediate results, a data structure called a table is used. The AprioriTid algorithm, unlike the Apriori algorithm, uses the transaction identifier (TID) to search for frequent itemsets, which significantly reduces the algorithm running time.

The FPG method uses a data structure called an FP-tree. It allows you to discover a frequent itemsets without candidate itemsets generation. Frequent itemsets are extracted directly from the FP-tree [10]. The FPG method is complicated in implementation but it gives a gain in the algorithm running time.

The ABBM method is based on a binary matrix [6, 9]. It is complicated to implement but it is more efficient in the algorithm running time than the FPG method. This is due to the matrix columns pruning, where there are candidates of frequent itemsets having support values less than the minimum support.

In [4], we analyzed the AprioriTid method for the formation of the indicators system that affects life quality.

In this paper, having compared the AprioriTid, FPG and ABBM methods to each other on such criteria as algorithm running time, useful and confident association rules mining, we propose to apply the ABBM method to solve this problem.

3 Basic metrics of association rules mining

There are many techniques which allow us solving the problem of association rules mining. They have the same mathematical approach but are different in their implementation methods. Let us consider the basic theoretical principles of these methods [14].

The association rule of context K is an expression of the form $A \to B$, where $A, B \subseteq M$.

The context K is a tuple (G, M, I), where G is a set of objects, M is a set of features, but $I \subseteq G \times M$.

When association rules are searched, special metrics are used: Support, Confidence, Lift.

Association rule $A \to B$ Support is a quantity defined by the formula:

$$\text{Support}(A \to B) = \frac{|(A \cup B)'|}{|G|} \tag{1}$$

The Support value indicates which part of the G objects contains $A \cup B$.

The Confidence of the association rules is defined by the formula:

$$\text{Confidence}(A \to B) = \frac{|(A \cup B)'|}{|A'|} \tag{2}$$

The Confidence value shows, which part of the objects that contain A, also contains $A \cup B$.

The following quantity is called the association rule utility (Lift):

$$\text{Lift}(A \to B) = \frac{|(A \cup B)'|}{|A'| \times |B'|} \tag{3}$$

In other words, the utility is the ratio of Confidence$(A \to B)$ to the Support(B). The Lift value indicates how useful the rule is. If the found utility value is more than 1, then the rule is considered to be useful.

The task of association rules mining is to find all association rules of the context for which the support and confidence values exceed the certain set values min_support and min_confidence, correspondingly.

First, the transaction database is scanned where transactions consisting of various elements are stored.

Then we carry out a search for frequent itemsets from 1-itemsets, 2-itemsets, 3-itemsets, etc. For this, such data structures, as a table, tree or binary matrix are used, depending on the method used.

The search of frequent itemsets is limited to the minimum support value of min_support, which is set by the user [2]. The association rule mining is performed in frequent itemsets and is limited to the value of the minimum confidence of min_confidence and utility (Lift). Usually the minimum confidence is set by every user.

4 Software of association rules mining

4.1 Software architecture development

To solve the problem of association rules mining, the special software was developed. Its architecture is shown in Fig. 1.

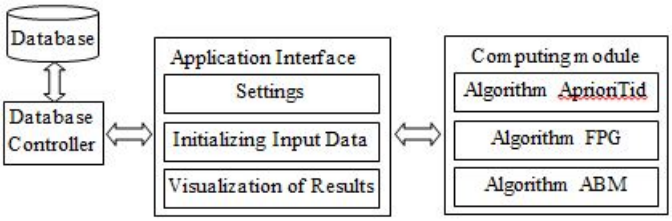

Fig. 1. Architecture of association rules mining software

The software consists of three modules: a user interface, computational module, database controller.

The user interface module is a set of dialog boxes that serves to provide interaction of the software modules and the user. When the software is started, the main window and the window for displaying the service information are opened. The main window consists of a control panel and a results panel. Using the control panel, the entire system of association rules mining is managed. The output pane consists of two windows: "Useful Rules", "Description".

The "Useful Rules" window displays all the useful and trustworthy rules. If you choose one of the rules there, then you can see the names of the indicators included in the associative rule and the values of its metrics, such as support, confidence and lift, in the "Description" window.

It should be noted that in the associative rule, all the indicators are presented in an encoded form, so the "Description" window allows you to analyze the indicators that are part of the rule.

The database controller is used to access the database. The database is represented as xml format files and consists of four tables: experiments, transactions, attributes (items), results. Input to the database is performed by importing some data from an Excel file. A transaction is a set of attributes that characterize a subject area under analysis, for example, "Lifetime". Each attribute is encoded with a unique integer value which identifies the former in a database.

The computing module is a software implementation of the AprioriTid, FPG, ABM algorithms for the association rules mining. To start calculations, you need to select one of the algorithms in the main program frame and specify its settings: minimum support (minsup), minimum confidence (minconf), utility (lift).

The program allows you saving the results in a database.

4.2 Association rules mining algorithms

In this paper, FPG and ABM the algorithms are presented. The AprioriTid algorithm was considered in [5].

FPG Algorithm

The basis of the Frequent Pattern-Growth method is the transformation of the transaction database into a tree structure, called the tree of frequent datasets (Frequent-Pattern Tree or FP-tree). The block diagram of the FPG algorithm is shown in Fig. 2.

The transaction database is scanned first, and frequent itemsets are selected. These are the items, the number of which is bigger and equal to the minimum support value in various transactions.

1. These frequent itemsets are sorted in descending order of their support values, for example, $(c, 5)$, $(b, 4)$, $(a, 3)$.
2. After sorting, the root node of the FP tree is created, which is denoted as ROOT.
   The tree construction begins with the enumeration of transactions.
   The tree is being constructed according the following rule. If for the next transaction tree element there is a node which name coincides with the element name, then the element does not create a new node, and the index of the corresponding node in the tree is increased by 1. Otherwise, a new node for this item is created and an index 1 is assigned to it.
3. Then you need to look through all the items in the loop and find all the paths in the tree that lead to the current item nodes. For each path, you have to calculate how many times the current item takes place in it and put it in a 2-tuple (set, index).
4. Next, remove the item itself (i.e. the set suffix) from the paths leading to it (i.e. the dial prefix).
5. Count how many times each item appears in the paths prefixes got in the previous step, and sort them in descending order of these values, getting a new set of transactions. It is based on the construction of a new FP-tree which is called a conditional FP-tree, since it is associated with only one object.
6. In this FP-tree, you need to find all the items (nodes) for which support (the number of occurrences in the tree) is equal to the minimum support or more. If the item occurs two or more times, then its indices (the occurrence frequencies in the conditional basis) are summed.

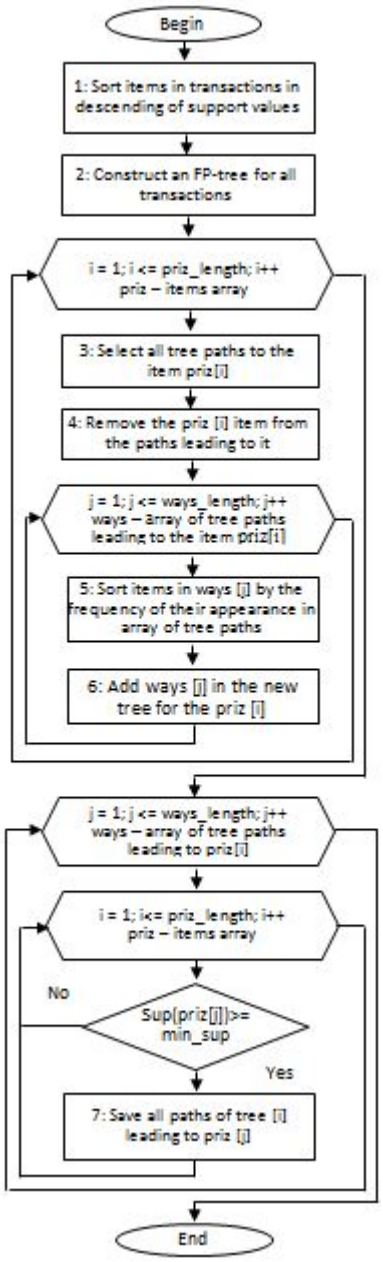

Fig. 2. Block diagram of the FPG algorithm

7. Starting from the tree root, you are to record the paths that lead to each node for which the support or index is bigger or equal to the minimum

support. Then you return the item (the template suffix) removed earlier and count the index or support obtained as a result. This result will be frequent itemsets.

ABBM Algorithm

An algorithm block diagram based on a binary matrix is shown in Fig. 3.

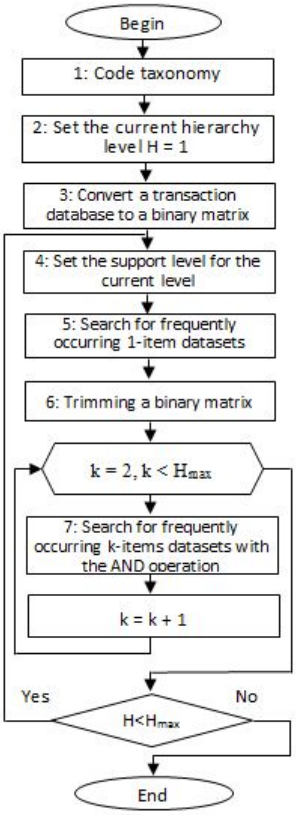

Fig. 3. Block diagram of the ABBM algorithm

The method essence is to mine multi-level association rules using the hierarchies' concept. The algorithm begins with the taxonomy coding. Then the first level of the H = 1 hierarchy is determined. Next, the transaction database is being converted to a binary matrix, and the minimum support value for this level is set. After that, we look for frequent 1-itemsets which are candidates for the rules. For each of them, support is calculated (i.e. the number of their repetitions in all the transactions involved in the experiment).

Next, the matrix columns are clipped, items support of which are less than the minimum support. Then 2-itemsets, 3-itemsets, ..., i-itemsets are generated in a loop by means of conjunction, where $2 \leq i \leq k$ and $k \leq \mathrm{H}_{\max}$.

For each of the following hierarchy level, the above stated steps (starting with step 4) are repeated. Once the maximum hierarchy level is reached, the algorithm is completed. Association rules are generated from the found of frequent itemsets.

5 The experiments results

To implement the association rules mining for indicators that affect the life quality, the system of indicators proposed by the authors in [12] was used. It was comprised of 46 indicators and attributes, including 8 indicators and 38 attributes which affect one or another indicator in different combinations. Five experiments were performed for which two indicators and seventeen attributes were selected. First, all the indicators and attributes of the system were encoded, then 30 transactions with the number of items from five to eleven were generated from the codes. For example, in the fifth experiment, the first transaction contains nine elements and has the following form: 91, 92, 93, 94, 95, 96, 97, 98, 99.

All experiments were carried out on a personal computer with the following characteristics: processor – AMD FX-6350 3.90GHz, 6 core; RAM – 8GB; Video card – NVIDIA GeForce GTX580 3GB; External memory – SSD Kingston 120GB; Operating system – Windows 10. The running time of the association rules mining algorithms was measured in ticks, where 1 tick is equal 1/10000000 seconds according to [1].

To determine the running time of the AprioriTid, FPG and ABBM algorithms, the association rules mining was carried out within the following parameters: min_support = 3; min_confidence = 0.8; lift > 1.

The graph in Fig. 4 shows how the algorithm running time and the number of transactions are correlated. The graph demonstrates that with 30 transactions the ABBM algorithm is executed 12 times faster than the AprioriTid algorithm and 1.6 times faster than the FPG algorithm.

Within the same parameters, a graph of relation between the rules number and the transactions number was constructed. It is presented in Fig. 5.

The graph shows that the number of rules generated with the AprioriTid method is 2.2 times more than that with the FPG and ABBM ones. It is due to the antimonotonicity property of the AprioriTid method. When you add new items to a transaction, all frequent transaction itemsets are saved. As a result, the number of new frequent itemsets increases thus the number of generated rules increases too.

For the next experiment, we set the min_support parameter equal to 10, the min_confidence and lift parameters remained the same as in the previous two experiments. The number of transactions was 20. It was done to generate a small number of rules.

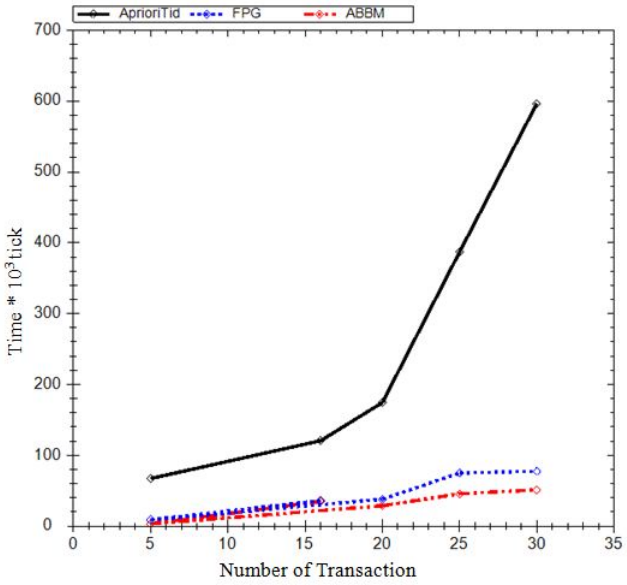

Fig. 4. Graph of relation between the algorithm running time and the number of transactions

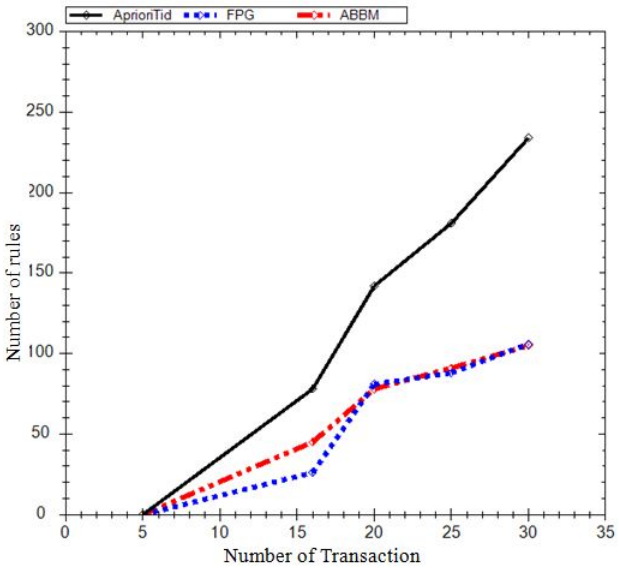

Fig. 5. Graph of relation between the number of rules and the number of transactions

The experiment resulted in obtaining five valid and useful association rules with the AprioriTid and ABBM methods and six rules by the FPG method. Since

the association rule is an implication, we united them by means of conjunction. Table 2 shows the results of the rule generation.

Table 2. Association Rules

Method	Association Rules
AprioriTid	$(92 \to 98) \wedge (94 \to 98) \wedge (98 \to 100) \wedge (98 \to 102) \wedge (97 \to 98)$
FPG	$(93 \to 94) \wedge (94 \to 97) \wedge (94 \to 102) \wedge (98 \to 94) \wedge (98 \to 97) \wedge (98 \to 102)$
ABBM	$(92 \to 98) \wedge (94 \to 98) \wedge (98 \to 100) \wedge (98 \to 102) \wedge (97 \to 98)$

The rules $(92 \to 98) \wedge (94 \to 98) \wedge (98 \to 100) \wedge (98 \to 102) \wedge (97 \to 98)$, obtained by the ABBM method, mean: (92 and 94 and 97) affect 98, and 98 in its turn affects (100 and 102).

The table analysis shows that the rules obtained with the AprioriTid and ABBM methods are the same and include six attributes: 92, 94, 97, 98, 100, 102, while the rules generated with the FPG method differ from them and include only five attributes: 93, 94, 97, 98, 102. They have four attributes in common: 94, 97, 98, 102. Since the two methods include the same attributes, they can be totally included in the life quality indicators system to further solve the problems of analyzing and forecasting. These indicators and attributes are as follows: 92 – Fiscal capacity per capita; 93 – Retail turnover per capita; 94 – Volume of paid services provided per capita; 97 – The growth rate of the minimum living wage; 98 – GRP per capita; 100 – The proportion of the population whose incomes are below the minimum living wage; 102 –Employment level.

6 Conclusion

Computational experiments were carried out with the developed software. They enabled us to obtain valid and useful association rules for the population life quality indicators. The correlation between the algorithms running time and the number of transactions was revealed. It was found that the ABBM method was the most efficient. The FPG method is 1.6 times slower than the ABBM method.

The correlation between the number of rules and the number of transactions was obtained. It demonstrates that the more transactions are present, the more rules are generated. As for the AprioriTid method it generates 2.1 times more rules than the ABBM and FPG methods because of its antimonotonicity property. This fact negatively influences on the AprioriTid algorithm running time.

The experiments results demonstrated that application of association rules mining methods will allow identifying the most significant indicators in any subject area. Further on we can use these indicators to solve various problems of analyzing and forecasting.

References

1. Electronic library of reference materials of Microsoft for the C# language. https://docs.microsoft.com/ru-ru/dotnet/csharp/language-reference
2. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I.: Fast discovery of association rules. In: Advances in Knowledge Discovery and Data Mining, pp. 307–328. American Association for Artificial Intelligence Menlo Park, CA, USA (1996)
3. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Proceedings of the Eleventh International Conference on Data Engineering. pp. 3–14. ICDE '95, IEEE Computer Society, Washington, DC, USA (1995)
4. Bilgaeva, L., Shirapov, D., Badmaev, G.: Formation of life quality indicators system through search algorithm of association rules. Proceedings of the Work-shop on CMDM 2016. Aachen, Germany, CEUR-WS 1726, 13–22 (2016)
5. Billig, V., Tsaregorodcev, N., Ivanova, O.: Building association rules in medical diagnosis. Programmnye produkty i sistemy: International journal 2, 146–157 (2016)
6. Chinta Someswara Rao, Ravi Babu, D., Shiva Shankar, R., Pradeep Kumar, V., Rajanikanth, J., Chandra Sekhar, C.: Mining association rules based on boolean algorithm – a study in large databases. International Journal of Machine Learning and Computing 3(4), 347–351 (2013)
7. Houtsma, M., Swami, A.: Set-oriented mining for association rules in relational databases. In: Proceedings of the Eleventh International Conference on Data Engineering. pp. 25–33. ICDE '95, IEEE Computer Society, Washington, DC, USA (1995)
8. Kostenchuk, M., Drozhzhin, N., Belousov, R.: Search of patterns in estimation of forest fire situation by weather conditions. Scientific and educational problems of civil protection 2, 61–66 (2014)
9. Liu, H., Wang, B.: An association rule mining algorithm based on a boolean matrix. Data Science Journal 6, 559–565 (2007)
10. Oreshkov, V.: FPG – an alternative search algorithm for association rules. https://basegroup.ru/community/articles/fpg (2014)
11. Pivovarova, N., Vidunova, S.: Data mining in the pharmaceutical business. Naukovedenie 8(6) (2016), http://naukovedenie.ru/PDF/166TVN616.pdf
12. Saktoev, V., Sadykova, E.: Sustainable Development of Regional Economic Systems with Environmental Regulations. ZAO "Economy", Moscow, Russia (2011)
13. Vercellis, C.: Business Intelligence: Data Mining and Optimization for Decision Making. Willey (2009)
14. Zayko, T.A., Oleinik, A.A., Subbotin, S.A.: Association rules in data mining. Vestnik NTU "KPI" 39(1012), 82–95 (2013)