Available online at www.worldscientificnews.com

World Scientific News

WSN 78 (2017) 131-142 EISSN 2392-2192

Application of the Apriori algorithm in the management of sports training

Monika Nawrocka[1,*], Mateusz Nawrocki[2], Karolina Kostorz[1], Agnieszka Dygacz[1]

[1]The Jerzy Kukuczka Academy of Physical Education in Katowice,
72A Mikolowska Street, 40-065 Katowice, Poland

[2]Institute of Archaeology, Faculty of Humanities, Maria Curie-Sklodowska University in Lublin,
5 Marii Sklodowskiej-Curie Street, 20-001 Lublin, Poland

E-mail address: monika.nawrocka@centrumstatystyczne.pl

ABSTRACT

The objective of the study was to test the Apriori algorithm of classification for obtaining training components that are connected with one another in the strongest way, in athletic training of race walkers. The presented algorithm expansion of the selection process will also generalize and obtain a system for recognizing the effectiveness of athletic training. Furthermore, particular properties of the criteria of component sets will be specified in terms of general selection in order to combine the relative and non-relative criteria.

***Keywords*:** management, sports training, algorithm, Apriori algorithm

1. INTRODUCTION

Data mining tools based on obtaining knowledge from data sets more often appear in analytics and optimization of sports training (АЖИППО et al. 2014; Percy 2015; Popovski et al. 2016; Yu 2015). A number of practitioners and coaches collect data and try to acquire information from such data allowing to obtain knowledge depending on the tested parameters that are decisive in the case of the sports result (Wiktorowicz et al. 2015). Some of the most advanced parameters relate to the psychomotor aspect, distribution of training loads, frequency of neural oscillations or muscle activity during workouts. Transactions or data is considered phenomena or processes in sports depending on the goal of the analyzed unexpected relations are subject to explorations covering various techniques, schemes or visualizations in such sets.

Considering the fact that data exploration is used for identification of association rules, an insight into the distribution of training loads may indicate which of them are associated. Such an identification of standards may be the basis for undertaking training decisions.

The Apriori algorithm was used to test the relation between athletic training components including the frequency of their occurrence in a 5-week athletic training. The studies aim at a quantitative assessment of relations between training indicators and at support of athletics training of master class race walkers.

2. THEORETICAL BACKGROUND

Management and control of sports training - the follow-up of a professional activity led by a coach, targeted at a specific sports goal, medium-term and long-term athlete's response to a given load (Ważny 1989, Sozański et al., 1999, Issurin 2010). These reactions are analyzed so that training effects can be attained (Lyakh et al., 2012). As a consequence, whether the planned training process is successful. Beyond the process of sports training for information on the stages of individual training, training cycles or training periods. Therefore, in the products presenting an example of the application of Apriori computational management to the training process, what gives better understanding, more specific planning and control of sports training.

2. 1. Method of data collection and analysis

The Apriori algorithm sets associative rules of learning in data or trans action sequences in databases (Aggarwal, Reddy 2013; Agrawal et al. 1996). Transactions or data is considered to be a set of components that is identified by an algorithm based on the threshold value of *C* frequencies being subsets of at least one C transaction/data (Argwal, Srikant 1994; Duarte et al. 2014). Particular items are expanded on to increasingly bigger sets of components. Nevertheless, these sets should often appear in the database. Apriori, based on “associating” frequently occurring component sets indicate tendencies in the database (Tsai, Chen 2004; Tan et al. 2006).

The problem of mining association rules defined by Agrawal (1996) and Tang et al. (2013) describes:

Let$I = \{i_1, i_2, \ldots, i_n\}$ will be *n* set of binary attributes called items. Let$D = \{i_1, i_2, \ldots, i_n\}$

be the set of transactions called database. Each transaction *D* features a unique transaction ID and contains a subset of elements in *I*. Rule is defined as $A \longrightarrow B$, where$A, B \subseteq I$ and$A \cap B = \emptyset$. A set of elements *A* and *B* is referred to as antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule respectively.

Databases used for step-by-step development of associative rule identification was obtained by recording the completed training units by ten master class race walkers during a 5-week athletic training. The participants represented the same sports level, i.e. master class.

In the case the sports level was lower than the required level, the participants would be excluded from the study.

Eight training units were entered to the databases, including: endurance range 1 (oxygen level ≥50%), endurance range 2 (oxygen level ≥70%), endurance range 3 (oxygen level ≥85%), endurance efficiency (special - progressive oxygen level), strength endurance, fast rhythm, long rhythm and athletic recovery. Each training element occurring in a given week was replaced and arranged into a set of components, and adjusted for each transaction in the database $D = \{i_1, i_2, \dots i\}$ (as shown in table 1.).

Table 1. Transaction Id data set, performed by a group of race walkers in the 5-week training.

Transaction ID (T)	Items
t1	{ER1, RL, ER2, ER1, ER1, RS, SER, ER1, BR}
t2	{ER1, ER1, RS, ER2, SER, ER1, ER1, SE, RS, ER1}
t3	{ER1, ER2, ER1, RS, E1, SE, RL, ER1}
t4	{ER1, RS, SER, ER1, ER1, RS, ER2}
t5	{BR, ER1, RS, ER1, ER2, ER3, ER1, ER1}

Legend: ER1- endurance range of 1, RS – rhythms short, RL – rhythms short, SER – special endurance, BR – biological regeneration, SE – strength endurance; Source: own elaboration

The WEKA software version 3.8 was used for classification, clustering and association analysis of the Apriori algorithm.

2. 2. Preprocessing

Base transaction data was stored in CSV format so that the Apriori algorithm could be launched. The threshold frequency was set prior to launching the algorithm.

Niu et al. (2016) think that there is a relation from point A to point B if the credibility is higher than the threshold value of the frequency. In this study, a threshold value of at least 60% was adopted for each single transaction for checking the number of repeated training components. This means that a given training element should occur at least three times in a 5-week training. The work of Roşca and Rădoiu (2015) was taken as an example. They indicated in their work that the time of the algorithm reduced as the frequency threshold increased.

Prior to starting calculations, implementation of the algorithm required to configure the Apriori implementation classes in Java code. The adopted Apriori class parameters are shown in the code below:

```
importjava.io.BufferedReader;
importjava.io.FileReader;
importjava.io.IOException;
importweka.associations.Apriori;
importweka.core.Instances;
publicclassMain
{
        publicstaticvoidmain(String[]args)
        {
                Instances data =null;

                try{
                                BufferedReaderreader=newBufferedReader(new
                FileReader( "...\\example2.arff" ));
                        data =newInstances(reader);
                        reader.close();
                        data.setClassIndex(data.numAttributes()- 1);
                }
                catch(IOException e ){
                e.printStackTrace();
                }
                doubledeltaValue= 0.05;
                doublelowerBoundMinSupportValue= 0.1;
                doubleminMetricValue= 0.5;
                intnumRulesValue= 20;
                doubleupperBoundMinSupportValue= 1.0;
                String resultapriori;
                Aprioriapriori=newApriori();
                apriori.setDelta(deltaValue);
        apriori.setLowerBoundMinSupport(lowerBoundMinSupportValue);
        apriori.setNumRules(numRulesValue);
        apriori.setUpperBoundMinSupport(upperBoundMinSupportValue);
        apriori.setMinMetric(minMetricValue);
                try
                        {apriori.buildAssociations( data );}
                        catch(Exception e ){
                                e.printStackTrace();}
                        resultapriori=apriori.toString();
                System.out.println(resultapriori);}}
```

Figure 1. The pseudo code of Apriori algorithms in WEKA.

The algorithm shown in Figure 1, was carried out in order to set Lower bound for minimum support of 0.1, rules with stores higher than value minimum metric score of 0.5 and number of rules to find of 20. Interactive reduction of the support was determined using a delta value at 0.05. The upper limit for the minimum support was 1.0.

3. ANALYSIS

Step 1: The Apriori algorithm sums up the number of transactions where each item occurs, i.e. occurrence of a given item in each of the five transactions.

Table 2. Numberpositions of items in transaction.

Item	Number of transaction
ER1	5
ER2	4
ER3	1
SER	3
RS	5
RL	2
SE	2
BR	2

Source: own elaboration

All possible components in transactions were generated [ER1], [ER2], [ER3], [SER], [RS], [RL], [SE], [BR]. The frequency for each single transaction was computed (how often a transaction component occurred in the database) $D = \{i_1, i_2, \dots i_5\}$. The frequency in all transactions was compared with frequency threshold.

Step 2: When analyzing the frequency of components in transactions as compared with the frequency threshold, four components were eliminated in transactions [ER3], [SE], [RL], [BR]. The remaining items repeated at least 3 times, were qualified.

Table 3. Number of classified individual items.

Item	Number of transaction
ER1	5
ER2	4
SER	3
RS	5

Source: own elaboration

Step 3: Possession of the repeating components with threshold frequency allows to create pairs of components starting from the first item. As a result of the above, it was no longer necessary to start from the second component which would provide the same effect as the one obtained when starting with the first component. The procedure consisted in combining pairs from the same first alphabet. This means generating two products on the basis of the item of the same alphabet (Duarte et al. 2014):

$$A \ \wedge B \ \rightarrow AB$$

$$A \ \wedge C \ \rightarrow AC \text{ and to bounded}$$

After entering all pairs and assigning components to letters of the alphabet, sets shown in Table 4 were obtained.

Table 4. Paris of 2-item transaction sets.

Item paris
ER1_ER2 ER1_SER ER1_RS ER2_SER ER2_RS SER_RS

Source: own elaboration

Scanning databases of transactions allows to obtain six 2-item transaction sets.

Step 4: The second step needs to be repeated relating to the set of 2-item transaction candidate list from Table 3 including the frequency of repeating pairs from the transaction databases contained in Table 1. The repeatability of those component pairs is shown in Table 4.

Table 5. Number of pairs of 2-item transaction sets.

Item pairs	**Number of transactions**
ER1_ER2	4
ER1_SER	3
ER1_RS	5
ER1_SE	3
ER2_SER	2

ER2_RS	4
ER2_SE	1
SER_RS	3
RS_SE	2

Source: own elaboration

The most often repeated 2-item set (5 times) was set [ER1_ER2]. Sets [ER2_SE], [RS_SE] and [ER2_SER] were repeated less often than the indicated threshold frequency.

Step 5: Based on the table from Step 4, pairs of the number of transactions repeated less than 3 times were excluded. The result of this procedure is shown in Table 6.

Table 6. Number of pairs classified 2-item transaction sets.

Item pairs	**Number of transactions**
ER1_ER2	4
ER1_SER	3
ER1_RS	5
ER1_SE	3
ER2_RS	4
SER_RS	3

Source: own elaboration

If you look at table 6, you can see pairs of components often executed among athletes.

Based on this knowledge, it is possible to proceed to the next step of the Apriori algorithm. It consists of searching a set of three components connected with one another, and then of supplementing the sets.

Step 6: Generating a 3-item set should be preceded with satisfaction of two pairs from the same first alphabet. This means that in a set of two items, AB, AC, AD and BC should be filled out in order to generate three products based on two items from the same alphabet:

$$AB \wedge AC \rightarrow ABC$$

$AC \wedge AD \rightarrow ACD$ and to bounded

Table 7. Paris of 3-item transaction sets.

Item pairs
ER1_ER2_RS RS_ER1_SER ER1_RS_SE ER1_ER2_SER ER1_RS_SER ER1_SE_SER

Source: own elaboration

Scanning databases of transactions allows to obtain six 3-item transaction sets.

Step 7: At this stage, the number of pairs of the 3-item component set are summed up.

Table 8. Number of pairs of 3-item transaction sets.

Item pairs	**Number of transactions**
RS_ER2_SER	2
ER1_RS_SE	2
ER1_ER2_SER	2
ER1_RS_SER	3
ER1_SE_SER	1
ER1_ER2_RS	4
ER1_ER2_SE	**1**

Source: own elaboration

The set of three training items often completed by participants at least three times is as follows: ER1_ER2_RS and ER1_RS_SER (Table 9).

Table 9. Number of pairs classified 3-item transaction sets.

Item pairs	**Number of transactions**
ER1_RS_SER	3
ER1_ER2_RS	4

Source: own elaboration

The ER1_ER2_RS set is often completed in athletic training of race walkers.

Table 10 shows the best rules of the association of component pairs which were assessed according to three indicators: confidence, lift, leverage and conviction. Confidence is calculated by dividing the probability of the items occurring together by the probability of the occurrence of the antecedent. In frequent item set, the confidences the ratio of the number of transactions of the type A,B and the total number of transactions containing item A, calculated as:

$$confidence(A \longrightarrow B) = \frac{supp(A \cup B)}{supp(A)}$$

$$supp(A \vee B) = \frac{|\{t \in T; X \subseteq t\}|}{|T|}$$

where, *support* means indicate the frequency of the selected set of database (Zhang et al. 2010).

The leverage is a ratio of the additional components. It provides information relating to improvement of the probability of the consequence of a given antecedent. It indicates the power of controlling a random co-existence of the antecedent based on their individual support and computation:

$$lift(A \longrightarrow B) = \frac{supp\ (A \cap B)}{supp(A) * supp(B)}$$

The conviction of a rule is measure of departure from independence, defined as:

$$conv(A \longrightarrow B) = \frac{1 - supp\ (A)}{1 - supp(A \rightarrow B)}$$

At first assumed specified minimum support of 0.1 and a user-specified minimum confidence of 0.9. It was computed to find All frequent item sets in a database and order to form rules in frequent item sets. Support and confidence are also the primary metrics for evaluating the quality of the rule generated by the model.

Table 10. Rulesfound in sports training 3-item transaction sets.

Set items paris	**Confidence**	**Lift**	**Leverage**	**Conviviton**
ER1_RS_SER	0.92	1.38	0.03	2.17
ER1_ER2_RS	0.91	1.36	0.03	1.83

Source: own elaboration

The quality of the associative rules of component sets indicates that among race walkers, 92% of athletes show endurance range of 1, short rhythms and special endurance. Just over 91% of athletes execute endurance range of 1, endurance range of 2 and short rhythms. The associative rule is considered to be interesting because it meets the minimum thresholds of support and credibility.

4. CONCLUSIONS

The objective of this paper was to support athletic training of race walkers based on the Mining Apriori algorithm. A study of data exploration allowed to obtain training components that are associated with one another in the strongest way and occur most frequently during athletic training. The Apriori algorithm also indicated a relation between endurance in the range between 1 and 2, as well as short rhythms. With the adopted lower frequency, relations of endurance in the range of 1, special endurance and short rhythms were observed. Nevertheless, there is a lack of works on indicating the associative rules of learning in sports data. The knowledge indicated in this work on learning associative rules of athletic training may support coaches or practitioners of sports in improving athletic skills, indicating components that are most frequently used with strong relations between one another.

The discussed management of the training process results from the interaction of the calculation tool and the training of the sport. Systematic control and manipulation of training volume parameters as well as training elements supports the training process. Combining coaching practice with the presented algorithm is crucial for the management, including planning and analysis of the training process, especially when the block division into training elements that enhance the effectiveness of sports preparation is made, thus making the training effects more controllable and predictable. (Counsilman and Counsilman, 1991).

The Apriori algorithm for sports data requires finding methods allowing for raster data scanning (Fujiwara et al. 2016). This is the basis for further research explaining the training load in relation in sports (Witten, Frank 2005). The problem of sports training items in cited other research authors was implemented in a different way than this work. Research method is innovative in optimizing the training of athletes and support management of sports training, allowing for very precise measurements choice of predictors.

References

[1] B. E. Counsilman, J. E. Counsilman, The residual effects of training. *Journal of Swimming Research*, (1991), 7 (1), 5-12

[2] B. Yu, Scientific research on track and field. *Journal of Sport and Health Science* (2015), 4(4), 307

[3] C. C. Aggarwal, C. K. Reddy, *Data clustering: Algorithms and applications, Chapman and Hall/CRC data mining and knowledge discovery series*. CRC Press: Boca Raton (2013).

[4] D. G. Roşca, D. Rădoiu, *Step-by-step model for the study of the apriori algorithm for predictive analysis*. Scientific Bulletin of the „ Petru Maior" University of Tîrgu Mureş 12 (2015), 44-47

[5] F. D. Percy, Strategy selection and outcome prediction in sport using dynamic learning for stochastic processes. *Journal of the Operational Research Society* (2015), 66(11), 1840-1849

[6] H. Mannila, *Methods and problems in data mining*. In: Proceedings of the International Conference on Database Theory, New York 1997, 41-55.

[7] H. Sozański (Polish), *Theory of sport training*. Sports Center: Warsaw, 1999.

[8] I. H. Witten, E. Frank, Data Mining, *Practical Machine Learning Tools and Techniques*. Morgan Kaufman Publishers: San Francisco, 2005.

[9] J. M. M. Duarte, A. L. N. Fred, F. Duarte, Constraint acquisition methods for data clustering. *Intelligent Data Analysis*, 18 (2014) 47-64

[10] K. Wiktorowicz, K. Przednowek, L. Lassota, T. Krzeszowski, Predictive Modeling in Race Walking. *Computational Intelligence and Neuroscience* 15 (2015).

[11] M. Fujiwara, Y. Kawasaki, H. Yamada, A Pharmacovigilance Approach for Post Marketing in Japan Using the Japanese Adverse Drug Event Report (JADER) Database and Association Analysis. *PLoS One* (2016), 11(4), 1-8

[12] P. N. Tan, M. Steinbach, V. Kumar, Introduction to data mining. Pearson Addison Wesley: Boston, 2006.

[13] P. S. M. Tsai, C. Chen: Mining interesting association rules from customer databases and transaction databases. *Information Systems*, 29 (2004), 685-696

[14] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. I. Verkamo, Fast discovery of association rules. In: Advances in Knowledge Discovery and Data Mining, Cambridge MA 1996, 307-328

[15] R. Agrawal, R. Srikant, Fast algorithms for mining association rules. In: Proceedings of the 1994 Very Large Data Bases Conference, Morgan–Kaufmann 1994, 487-499.

[16] S. Sarawagi, S. Thomas, R. Agrawal: Integrating association rule mining with relational database systems: Alternatives and implications. In: Proceedings of the ACM SIGMOD *International Conference of Management of Data*, Seattle 1998, 343-354.

[17] V. Issurin, Blokowaja periodizacja sportiwnoj trenirowki. Sowietskij Sport, Izdatelstwo: Moscow (2010).

[18] V. Lyakh, P. Bujas, L. Gargula, Training effects in the preparation of highly qualified football players: Review. *Anthropomotorics* (2012), 59, 121-135.

[19] W. Nengsih, A comparative study on market basket analysis and apriori association technique. In: *Conference proceeding of 3rd International Conference on Information and Communication Technology*, Bali 2015, 461-464

[20] Z. Niu, Y. Nie, Q. Zhou, L. Z. J. Wei, A brain-region-based meta-analysis method utilizing the Apriori algorithm. *BMC Neuroscience* (2016), 17-23.

[21] Z. T. Popovski, T. Nestorovski, M. Wick, A. Tufekchievski, A. Aceski, S. Gjorgjievski, Molecular-genetic predictions in selection of sport talents and ethical aspect of their application. *Research in Physical Education, Sport & Health* (2016), 5(1), 57-63.

[22] Z. Ważny (Polish), Model performance indicators for the championship. RCMSzKFiS: Warsaw (1989).

[23] Z. Zhang, H. Cheng, X. Chu, Aided analysis for quality function deployment with an Apriori-based data mining approach. *International Journal of Computer Integrated Manufacturing*, 23 (2010) 673-686

[24] А. Ю. АЖИППО, Н. А. БАЛОНИН, В. А ДРУЗЬ, В. С. СУЗДАЛЬ, Finite system optimization sports equipment movements. *Slobozhanskyi R & Sports Bulletin* (2015), 46(2), 1-9

(Received 14 June 2017; accepted 11 July 2017)