ORIGINAL

PERFORMANCE INDICATOR SELECTION USING DECISION TREES IN ELITE HANDBALL

SELECCIÓN DE INDICADORES DE ÉXITO EN BALONMANO DE ÉLITE A TRAVÉS DE ÁRBOLES DE DECISIÓN

Cabrera Quercini, I.1; González-Ramírez, A.1; García Tormo, J.V.2; Martínez, I.3

1 Profesor titular. Instituto Universitario Asociación Cristiana de Jóvenes (Uruguay). icabrera@iuacj.edu.uy, agonzalez@iuacj.edu.uy
2 Profesor contratado doctor. Departamento de Educación Física y Deportiva. Universidad de León (España). jvgart@unileon.es
3 Profesor Titular. Departamento de Educación Física y Deportiva. Universidad de León (España). imarm@unileon.es

Spanish-English translator: External service www.proofreadingservices.com

Código UNESCO/UNESCO code: 5899 Otras especialidades pedagógicas (Educación Física y Deporte).

Recibido 21 de septiembre de 2020 Received September 21, 2020
Aceptado 29 de mayo de 2021 Accepted May 29, 2021

ABSTRACT

The aim was to analyze the performance indicators in the European Men's Handball Championship using decision trees as artificial intelligence models. The observational methodology was used. The sample was composed of 87 matches from the 2016 and 2018 Men's European Handball National Championships. As the most important result, the model identified three relevant variables to achieve high precision to predict handball results. In conclusion, the use of these models allow to greatly reduce the complexity in the analysis of the performance indicators in handball.

KEYWORDS: handball, performance analysis, prediction, artificial intelligence, decision trees, performance indicators.
RESUMEN

El objetivo fue analizar los indicadores de éxito en el Campeonato Europeo de balonmano masculino utilizando árboles de decisión como modelos de inteligencia artificial. Se utilizó la metodología observacional. La muestra fue compuesta por 87 partidos de los Campeonatos de Europa masculinos de selecciones de balonmano 2016 y 2018. Como resultado más importante, el modelo identificó tres variables relevantes para alcanzar una precisión elevada en la predicción de resultados de balonmano. Se concluye que la utilización de estos modelos permite reducir ampliamente la complejidad en el análisis de los indicadores de éxito en balonmano.

PALABRAS CLAVE: balonmano, rendimiento, predicción, inteligencia artificial, árboles de decisión, indicadores de rendimiento.
INTRODUCTION

The identification of performance indicators in handball as a line of research has developed rapidly in recent decades and has provided a great deal of useful information for coaches. According to Srhoj, Rogulj, and Katić (2001), the result of a match is the product of the interaction of the two participating teams manifesting itself through the game’s elements and external environmental influences. Those elements that have the most influence on the result are identified as performance indicators. They are “a selection or combination of action variables that tend to define some or all aspects of performance” (p. 739). They are ordinarily used by coaches to evaluate individual or team performances and sometimes to compare these with the opponent or groups of players (Hughes & Bartlett, 2002; O'Donoghue, 2014).

In the last decades, attempts have been made to identify the best performance indicators in handball. These have led to some agreement on the importance of certain variables that allow winning and losing teams to be distinguished (Saavedra et al., 2017; Beiztegui-Casado et al., 2019). Fast breaks, front-line shots, and goalkeeper involvement are the performance indicators that appear most frequently in the studies reviewed. Srhoj et al. (2001) state that fast breaks contribute the most goals along with front-line shots and penetrations, as well as being the most effective shot. Rogulj, Srhoj and Srhoj (2004), Grujić, Vuleta, and Milanović (2006), Saez, Roldán, and Feu (2009), Hernández et al. (2010), Foretić, Rogulj, and Trminić (2010), Gutiérrez Aguilar (2011), Bilge (2012), Hassan (2014), and Amatria et al. (2020) agree that the indicators of fast break success are highly influential in determining match outcomes.

On the other hand, first-line shots are indicators of success both from an offensive and defensive perspective (i.e., goalkeeper saves). Although more distant shots are less effective (Srhoj et al., 2001), good offensive records from this zone are a positive measure for determining winning teams (Bilge, 2012; Ferrari, dos Santos, & Vaz, 2014; Grujić et al., 2006). Conversely, lower shooting efficiency in this zone is characteristic of losing teams (Foretić et al., 2010; Gutiérrez Aguilar, 2011).

According to Pascual, Lago, and Casais (2010), goalkeeper efficiency and shot efficiency are performance indicators that are related to a higher probability of winning. Winning teams had a higher number of goalkeeper saves than losing teams (Daza, Andrés, & Tarragó, 2017). Furthermore, goalkeeper effectiveness is associated with the team's final ranking in the tournament (Hansen et al., 2017). Saez et al. (2009) argue that goalkeeper saves of 6-metre shots characterise winning teams. The goalkeeper is present in all performance indicators presented previously, since, with an effective goalkeeper, the efficiency rates of shots decrease. The importance of the goalkeeper is therefore clear.

In recent years, new forms of analysis of performance indicators have been introduced. These are based mainly on systematic observation (Anguera & Hernández Mendo, 2015) and are called notational analysis in the field of sports science (Gómez-Ruano, 2017). In addition, decision trees (DTs) are the most...
common and powerful data analysis and prediction structures in artificial intelligence (AI). Their use involves a low computational cost, which translates into faster results and helps machine learning. The biggest benefit compared with other AI models is that the findings are easily explained, because its tree format shows the classification path through branches (Marsland, 2015). This model falls under supervised learning in its two modalities: regression and classification.

One advantage of decision trees is that they apply feature selection as part of the training process, making it a very efficient model (Guyon & Elisseeff, 2003). However, their predictions are not very accurate compared with other AI models. They are also unstable, as small changes in the input data can have a major impact on the structure of the tree and generate errors in the initial part of the tree that are transferred to the rest of the branches (Murphy, 2012). Ben-David and Shalev-Shwartz (2014) state that these algorithms generally return trees that are too long and complex, which does not help their implementation. To solve this problem, alternatives can be applied, for example lowering the number of iterations, increasing the minimum number of records required to split the child branches, or performing pruning after the tree has been created.

The present study aimed to analyse performance indicators in the European Men's Handball Championship using decision trees as artificial intelligence models.

MATERIAL AND METHODS

We followed the parameters established by observational methodology (Anguera & Hernández Mendo, 2015) to achieve an objective observation that guaranteed the quality of the data. The observational design proposed was I/P/M: ideographic, punctual, and multidimensional (Anguera, Blanco, Hernández Mendo, & López, 2011).

Sample

Following the sampling levels proposed by Anguera and Hernández Mendo (2013), the first level of inter-sessional sampling consisted of 87 of the 95 matches of the 2016 and 2018 Men's European Handball National Team Championships. Matches that ended in a draw were excluded.

The second level of sampling corresponding to the intra-sessional sample comprised 174 information vectors, one per team and match, with the records provided by the European Handball Federation (EHF) through its website. These records are published after each match according to frequency and order. For the present study, only the frequency type data were taken.

Variables and procedure

To generate the database, the final total statistics for each team per game were collected, and all the records were combined in a single Microsoft Excel file for
subsequent analysis. The information was distributed across a total of 74 variables grouped according to the following macro-criteria: 1) Identification 2) Offensive – shooting efficiency 3) Punishments – the quality of play 4) Attacking efficiency 5) Goalkeeper efficiency.

**Table 1.** Predictor variables included in the decision trees

<table>
<thead>
<tr>
<th>Offensive</th>
<th>Quality of play</th>
<th>Attacking efficiency</th>
<th>Goalkeeper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goals</td>
<td>Goals</td>
<td>Yellow Card</td>
<td>Attacks</td>
</tr>
<tr>
<td>Shots</td>
<td>Shots</td>
<td>Red Card</td>
<td>Attack Efficiency</td>
</tr>
<tr>
<td>Shot Efficiency</td>
<td>%tanz</td>
<td>2 min. Suspension</td>
<td>Player Majority Goals</td>
</tr>
<tr>
<td>7m Goals</td>
<td>7mPG</td>
<td>2 + 2 min. Suspension</td>
<td>Player Majority Attacks</td>
</tr>
<tr>
<td>7m Shots</td>
<td>7mPS</td>
<td>Assists</td>
<td>Player Majority Attacks</td>
</tr>
<tr>
<td>7m Efficiency</td>
<td>7m%</td>
<td>Received 7-metre Fouls</td>
<td>Player Minority Goals</td>
</tr>
<tr>
<td>6m Goals</td>
<td>6mCG</td>
<td>Turnover</td>
<td>Player Minority Attacks</td>
</tr>
<tr>
<td>6m Shots</td>
<td>6mCS</td>
<td>Technical Faults</td>
<td>Player Minority Efficiency</td>
</tr>
<tr>
<td>Wing Goals</td>
<td>WingGoals</td>
<td>Steals</td>
<td>Positional Attacks</td>
</tr>
<tr>
<td>Wing Shots</td>
<td>WingShots</td>
<td>Blocked Shots</td>
<td>Positional Attacks Efficiency</td>
</tr>
<tr>
<td>Breakthroughs Goals</td>
<td>BTG</td>
<td>Penalty 7-metre Fouls</td>
<td>Fast Breaks</td>
</tr>
<tr>
<td>Breakthroughs Shots</td>
<td>BTS</td>
<td>Total Turnover</td>
<td>Fast Break Efficiency</td>
</tr>
<tr>
<td>Fast Break Goals</td>
<td>FBG</td>
<td>Turnover proportion</td>
<td>Individual Fast Break Goals</td>
</tr>
<tr>
<td>Fast Break Shots</td>
<td>FBS</td>
<td></td>
<td>Individual Fast Break Goals</td>
</tr>
<tr>
<td>Fast Throw off Goals</td>
<td>FTOG</td>
<td></td>
<td>Individual Fast Break Efficiency</td>
</tr>
<tr>
<td>Fast Throw off Shots</td>
<td>FTOS</td>
<td></td>
<td>Team Fast Break Goals</td>
</tr>
<tr>
<td>9m Goals</td>
<td>9mG</td>
<td></td>
<td>Team Fast Break Goals</td>
</tr>
<tr>
<td>9m Shots</td>
<td>9mS</td>
<td></td>
<td>Team Fast Break Efficiency</td>
</tr>
</tbody>
</table>

Through AI analysis, decision tree models were constructed in IBM SPSS Modeler 18 software. The database partition was 70% for the training set (119 records) and 30% for testing (55 records). The data in the training set allowed the algorithm to train, while the test set corresponded to the data that the model did not know and tried to predict to determine the final performance of the created model. Of the 74 variables available, identification variables were excluded, with 67 finally forming part of the group of predictor variables. The algorithm used was C5.0, and it was asked to favour generalisation in exchange for the accuracy of the training set.
As part of the process of elaboration, the minimum value requested in the daughter branches of the decision tree was modified. In this way, the depth and complexity of the tree changed, starting with a minimum value of 2 (DT2), 4 (DT4); and finally 10 (DT10), generating three different decision trees. The final and best performing model was asked to produce, in addition to the decision tree, a set of rules to assist in the description of the classification.

RESULTS

As the minimum log for the splitting of the daughter branches increased, there was a decrease in the accuracy of the training set and a slight increase in the test set. The last decision tree emerged as the most accurate tree because it made a better prediction for the set of matches with the unknown outcome (Table 2).

<table>
<thead>
<tr>
<th></th>
<th>DT2</th>
<th>DT4</th>
<th>DT10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>99.16%</td>
<td>95.80%</td>
<td>89.08%</td>
</tr>
<tr>
<td>Test set</td>
<td>74.55%</td>
<td>74.55%</td>
<td>76.36%</td>
</tr>
</tbody>
</table>

This decision tree was configured with a minimum of 10 records per daughter branch (DT10). This determined a tree with a depth of 3 levels and 6 nodes (Figure 2). Three variables were selected by the model for the branching of the tree. They were (in order of importance): goalkeeper efficiency; shooting efficiency; and attack efficiency (Figure 1).

The accuracy of DT10 reached 89.08% in the training set and 76.36% in the test set. The performance was 0.92 and 0.736 AUC (area under the curve), respectively.
Figure 2. DT4 decision tree configuration. Winners (GANA) and losers (PERD)

The cut-off points for branching were 53% for shooting efficiency, 32% for goalkeeper efficiency, and 58% for attacking efficiency. This decision tree presented the best performance because of its accuracy and simplicity. Therefore, it was asked for the rules that governed the classification of winners and losers. There were four:

- Team wins if $\rightarrow$ attacking efficiency $>$ 58.
- Team wins if $\rightarrow$ shooting efficiency $>$ 53 and goalkeeper efficiency $>$ 32
- Team loses if $\rightarrow$ shot efficiency $\leq$ 53
- Team loses if $\rightarrow$ attacking efficiency $\leq$ 58 and goalkeeper efficiency $\leq$ 32

DISCUSSION

By analysing the performance indicators in the European Men's Handball Championship using decision trees as artificial intelligence models, it was possible to reduce the set of indicators to just three variables and to rank their importance with a high degree of accuracy:

1. Goalkeeper efficiency
2. Shooting efficiency
3. Attacking efficiency

That goalkeeper performance was the main classifier between winning and losing teams makes sense; its relevance has been previously identified in important tournaments such as IHF World Cups (Daza et al., 2017; Hansen et al., 2017) and the Pan-American (Cabrera & González, 2015; González, Bermúdez, Martínez, & Chirosa, 2017) and Spanish national tournaments (Pascual et al., 2010; Saez et al., 2009).

In terms of shot efficiency, significant differences in favour of winning teams were found by Saez et al. (2009), Cabrera and González (2015), Ferrari et al. (2014), Hassan (2014), and Saavedra, Borgeirsson, Kristjánsdóttir, Chang, and Halldórsson (2017). Furthermore, losing teams were directly associated with missed shots (Cabrera & González, 2015; Daza et al., 2017), and there were significant differences according to tournament rankings (Noutsos, Rousanoglou, Meletakos, Bayios, & Boudolos, 2018).
Finally, and in direct relation to the previous two indicators, attacking efficiency was an aspect of the game that differentiated winning and losing teams. It grouped together different elements of finishing, such as shooting, and quality of play (e.g., turnovers). It also differentiated winners and losers in studies of the Copa del Rey 2008 (Saez et al., 2009), the ODESUR, and Pan-American 2014 (Cabrera & González, 2015; González, Botejara, Martínez, & Chirosa, 2016) tournaments.

Using just these three variables, this model was able to predict more than three-quarters of the outcomes. Although it did not achieve exact precision, its advantage lay in the fact that it did not require too many human or computational resources for data recording and processing. Furthermore, it required fewer predictors than all the studies consulted on outcome prediction in sports that have used decision trees: 28 in Delen, Cogdell, and Kasap (2012), 15 in Soto Valero (2016), 8 in Thabtah, Zhang, and Abdelhamid (2019), and 4 in Joseph, Fenton, and Neil (2006).

The process to achieve the highest accuracy involved modifying the minimum records per daughter branch parameter in each decision tree. The number of predictors was reduced and the accuracy for the training set decreased, though the accuracy of the test set increased. According to Ben-David and Shalev-Shwartz (2014), this is an advantage, because the accuracy of the test set did not decrease even when the tree was simplified.

There are considerably fewer machine learning sports studies using decision trees than those using models such as artificial neural networks and support vector machines, although precedents can be found in football and basketball, in particular the NBA (Bunker & Susnjak, 2019).

The DT10 model was more accurate than the Naïve Bayes algorithm used by Joseph et al. (2006) to predict the outcome of English Premier League matches (45.77%) and the decision tree applied by Soto Valero (2016) to predict MLB results (58.62%). However, it was less accurate (83%) in predicting NBA results compared with a model that combined linear regression with a decision tree (Thabtah et al., 2019), and below 86% in predicting NCAA American football games using a decision tree (Delen et al., 2012).

Sports analysts who identify a small number of relevant variables for performance purposes will be able to focus on particular elements of the game and not on others, depending on the tournament or team being studied. The work of researchers will be to look for answers (with coaches and players) quickly and accurately using certain indicators (Gómez-Ruano, 2017) and use the knowledge gained during competitive matches and/or in training. The results of the present study were limited to elite European men's handball, so further research in other national and international contexts is needed.
CONCLUSIONS

The three main performance indicators that were identified using decision trees allowed us to achieve a good performance in results prediction. The main indicator identified was goalkeeper efficiency, followed by shooting and attacking efficiency.

The use of decision trees as a machine learning tool for the identification of performance indicators in handball has proven to be both functional and very useful. The present study has verified that the problem can be simplified by determining the most important variables. More data are required to further our knowledge of handball tournaments and improve the predictive capacity of the model.
REFERENCES


_Número de citas totales / Total references: 36 (100%)_  
_Numero de citas propias de la revista / Journal’s own references: 3 (8,3%)_