From Tree to Forest: Determining the Probability of Scoring a Goal in Football Games

Jan Hendrik Krone and Johannes Fischer

Technische Universität Dortmund, 44227 Dortmund, Germany
hendrik.krone@tu-dortmund.de
johannes.fischer@cs.tu-dortmund.de

Abstract. More and more teaching tools and materials on the topic of decision trees are being developed, since they play a central role in the field of artificial intelligence. We developed teaching materials that are based on existing ones, which we extend with the Random Forest algorithm [1]. This has the advantage that the algorithm can be embedded in a group work. We used the calculation of goal probabilities in football as an example, which many students are familiar with from television and video games.

Keywords: Decision Tree · Random Forest · Machine Learning · Football · Soccer · Lower Secondary School

1 Introduction

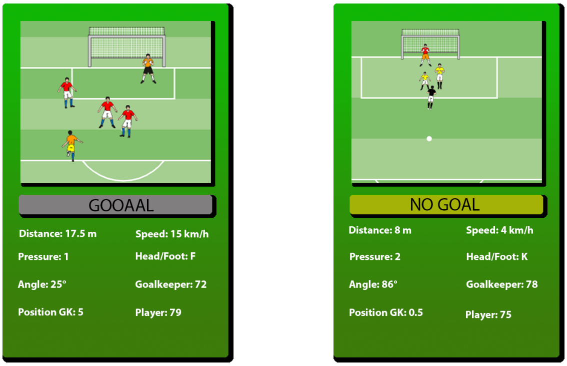

Fig. 1: Example of two playing cards. On the left for a goal and on the right for no goal.

We present a learning approach using playing cards (see fig. 1)[1] and videos of real football scenes. The general motivating approach is to let the students

[1] The material can be downloaded at the following link: https://tu-dortmund.sciebo.de/s/xtkh1ZMtQ2kFK1X

J.-P. Pellet and G. Parriaux (Eds.): ISSEP 2023 Local Proceedings, pp. 259–262, 2023.
https://doi.org/10.5281/zenodo.8432037

watch football videos, stop the video just before the striker's shot and use the *Random Forest* algorithm to predict the outcome, which is verified afterwards by continuing the video. Decision trees have been previously learned with the ID3 algorithm [4], using the playing cards containing training data from real football scenes. The selection of a context from the students' everyday life brings several advantages, such as motivation and interest. Additionally, a context-based approach makes it easier to understand abstract concepts such as algorithms [3].

2 Theoretical Background

Expected Goals: For the English Premier League, the statistics company OPTA has published a first model of expected goals in 2012 to calculate goal probabilities. For our lesson, we use the features *distance* (0-100), *position goalkeeper* (0-5), *pressure* (0-11), *player strength* (0-100), *goalkeeper strength* (0-100), *angle* (1-179 degrees), *speed* (0 km/h-35 km/h), and the binary feature *head/foot* (see fig. 2).

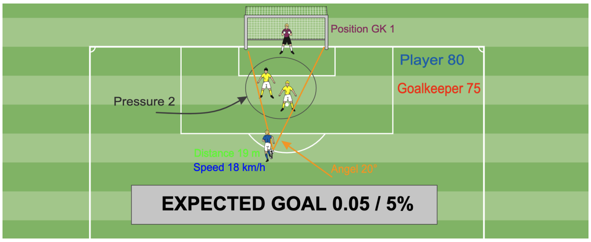

Fig. 2: Visualization of features that are relevant for the goal probability.

ID3 and Random Forest: The teaching unit is based on the *ID3* (Iterative Dichometer 3) algorithm to construct a decision tree. Then the *Random Forest* algorithm combines multiple trees.

ID3 is an algorithm to construct a simple decision tree [4]. The main idea is to choose a random training set and form a decision tree that correctly classifies all objects. To form a decision tree, one has to choose a feature for the root of the tree. All features are tested for their information content, and the best feature is picked as the root. Then the process is repeated with each new node.

The idea of the Random Forest algorithm [1] is to create multiple trees and then make a prediction by a majority vote. When initializing each tree, random features are used for its creation [2].

3 Implementation in school

3.1 One Feature / One Dimension

The first lesson starts with some football scenes to motivate the students. The teacher stops the video just before a player shoots. The students are asked about the outcome of the game scene and for their reasoning. The features from the explanations are collected and discussed. This results in the set of features, which is present on the cards (see section 1).

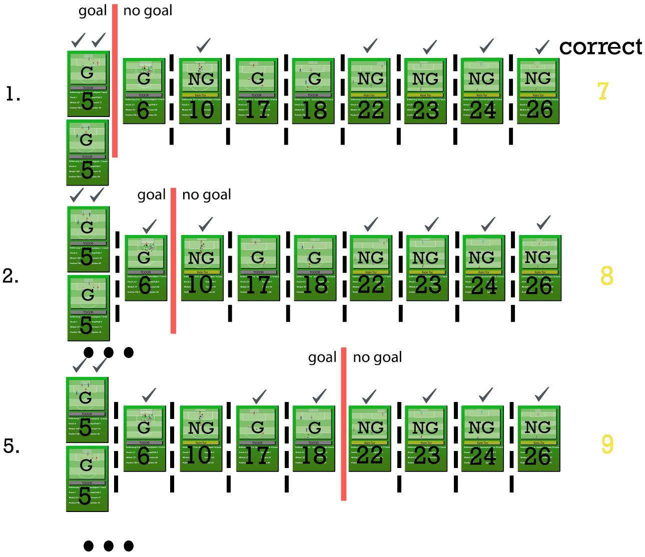

Fig. 3: Test of the feature *distance* (number on the card) for the information content. Correctly classified cards are indicated by check marks. The best dividing line is between distances 18 and 22.

When selecting the first feature in the decision tree, the first step is to find the best feature. How this process is carried out in the lesson is shown in fig. 3. The students each draw five cards from the *goal* deck and five cards from the *no goal* deck. Each group sort different cards features, seeking optimal dividing lines by exploring all options, counting correct classifications. They repeat this for all positions until all dividing lines are tested. Finally, the students compare their solutions and the best feature can be determined. This provides an opportunity to talk about the concepts of *overfitting* and *underfitting* for the first time.

3.2 Two Features / Two Dimensions

To counteract the limitation of a single feature, multiple features are used. For the extension, the students need to be able to represent two dimensions. They create a coordinate system to plot the data. Each student gets two random features, which are used for the axes. With the help of the two-dimensional representation, it is now relatively easy to construct the decision tree (see fig. 4).

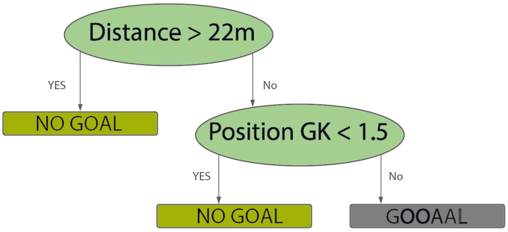

Fig. 4: Example of a possible decision tree with two features.

3.3 Random Forest

In the third part we try to improve the method by using predictions with a probability, instead of classifications. For this we use a group process to analyze the goal scenes. The decision trees created in the previous part are used again. All decision trees are evaluated simultaneously and the results are collected on the board. This collection forms a probability distribution. This method helps to predict a lot of game scenes correctly. Other scenes where goals are scored from far away or where strikers look particularly unlucky cannot be predicted correctly.

4 Future work

The units have been tested several times with different groups of lower secondary students. The students and the teachers consistently gave positive feedback. Teaching ML with the context of football is a suitable method. Next we will evaluate the effectiveness of the materials on students.

References

1. Breiman, L.: Random forests. Machine Learning **45**, 5–32 (2001)
2. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer New York (2009)
3. Nijenhuis-Voogt, J., Bayram-Jacobs, D., Meijer, P.C., Barendsen, E.: Omnipresent yet elusive: Teachers' views on contexts for teaching algorithms in secondary education. Computer Science Education **31**(1), 30–59 (2020)
4. Quinlan, J.R.: Induction of decision trees. Machine Learning **1**(1), 81–106 (1986)