DataDebugging: Enhancing trust in soccer action-value models by contextualization

Maaike Van Roy[1,2][0000-0001-8959-3575] and Jesse Davis[1,2][0000-0002-3748-9263]

[1] Department of Computer Science, KU Leuven, Leuven, Belgium
[2] Leuven.AI
{firstname.lastname}@kuleuven.be

Abstract. In sports analytics, tree ensembles are used to tackle a variety of tasks. In soccer, tree ensembles are commonly used to predict the expected values of players' actions. While these models perform well in general, there are two scenarios where their predictions should be treated with caution. First, if the data contains annotation errors, then the model's prediction is inherently wrong. Second, and more subtly, are actions that are highly dissimilar to what was seen during training. Machine-learned models struggle with extrapolation, and hence the model's value for such actions may be unreliable. This work aims to automatically flag the above two scenarios to help contextualize such models' predictions.

Keywords: Soccer analytics, Data debugging, Valuing actions, Trust.

1 Introduction

Tree ensembles (e.g., random forests, gradient boosted trees) are very popular as they achieve state-of-the-art performance on many tasks. In sports analytics, they have been used to solve tasks ranging from match prediction [1,2] to injury and fatigue prediction [3,4]. Specifically, in soccer, tree ensembles form the basis of many approaches for valuing player's actions (e.g., shots, passes, dribbles) [5,6,7,8,9]. This has various downstream use cases such as player analysis and recruitment, and match preparation.

While such action-value models perform quite well, there are two situations in which the predicted values cannot always be trusted. First, these models are applied to data that is annotated by humans while watching videos of matches. These annotators sometimes make mistakes. When applied to incorrectly annotated data, the model's predictions will be based on incorrect information and will inherently be wrong. Second, the model's predictions should be treated with caution when it is confronted with actions that are highly dissimilar to what the model has seen during training because it is well known that learned models struggle to extrapolate beyond the training data [10]. This paper discusses how to automatically detect these potentially abnormal situations by computing how dissimilar they are to examples the model was trained on. The flagged abnormal examples can then be corrected or removed before moving on to the downstream use cases. We evaluate our approach on two public event stream data sets and show that it can detect annotation errors and rare situations (e.g., extraordinary shots).

2 Data

We use publicly available event stream data from the UEFA EURO 2020 and the FA Women's Super League 2020/21.[1] This data captures information about all on-the-ball actions during games (e.g., type, start and end location, result, executing player). Each data set is split into a train set used to learn the action-value models, and a test set in which abnormal examples will be detected. For each data set, we use an 80-20 game-wise split. This yields approximately 90,000 and 200,000 train examples, and 20,000 and 50,000 test examples, for the respective data sets.

3 Detecting abnormal examples in action-value models

Detecting if an unseen example is abnormal requires two things: (1) a trained action-value model which can be queried, and (2) an approach to score the given example based on how abnormal it is with respect to the model. Next, we describe both parts.

3.1 Training the action-value model

We use a simple tree-based action-value model that, given a sequence of the three latest actions, predicts the probability that the team currently possessing the ball will score a goal in the next 10 actions [5]. We use the same features as in the original work to describe a sequence. During training, we use a 5-fold cross validation grid search to optimize the depth (i.e., 3, 5, 8) and the number of trees included (i.e., 50, 75, 100).

3.2 Detecting abnormal examples in tree ensembles

We use Devos et al.'s [11] OC Score approach that is designed to identify dissimilar test examples when using tree ensembles. Instead of using an example's feature representation, it encodes each test example as the ordered set of leaf nodes that are reached when executing the ensemble, which is called the example's output configuration (OC). Even though two examples may seem similar when comparing their feature representations, small changes in these features can yield very different output configurations and hence final predictions. By comparing the output configuration of a test example to those of the examples the ensemble was trained on, the approach can detect dissimilar test examples for which the ensemble's prediction may not be trusted:

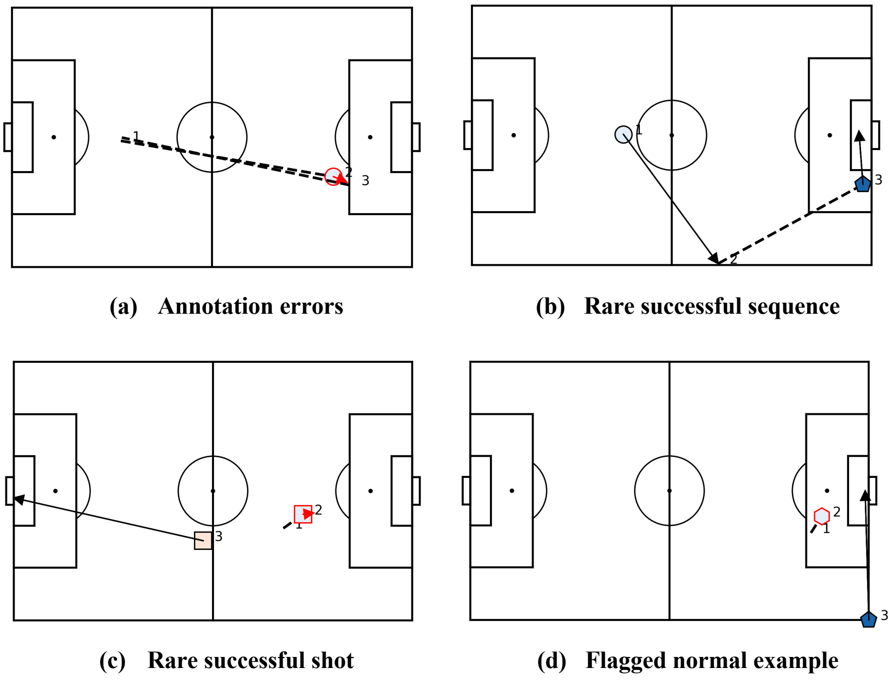
$$\text{OC-score}(x) = \min\{\text{Hamming}(\text{OC}(x), \text{OC}(x')) \mid x' \in R\}$$

where R is a reference set containing all correctly classified training examples with the same label that is predicted for the test example x. For each data set, we apply this method and flag a test example as abnormal when it's OC-score is greater than the average training set OC-score plus one standard deviation.

[1] https://github.com/statsbomb/open-data

4 Results and discussion

To validate whether our approach can detect both annotation errors and rare examples, we compare the flagged examples with their corresponding video. However, it's not straightforward to obtain videos of each example. Therefore, for each data set, we selected 10 examples out of the top 100 flagged abnormal test examples for which videos could be obtained. Based on the videos, two out of the 20 flagged sequences were found to be annotation errors, 11 could be considered infrequent ones, and 7 were found to be rather normal sequences. Fig. 1 shows four representative examples identified by the approach. First, it identifies the incorrectly annotated dribbles of Faye Bryson and Caroline Weir during the Bristol - Manchester City match (Fig. 1a). Second, it identifies infrequent sequences such as the well-executed goal-sequence starting with a long high pass by Italy against Switzerland during the EURO 2020 (Fig. 1b). Third, it picks up on exceptional goals that rarely occur, such as Patrick Schick's goal from the halfway line during the EURO 2020 (Fig. 1c). Fourth, it identifies the sequence involving the corner-kick of Katie Zelem during the Manchester United - Aston Villa match as abnormal (Fig. 1d). However, from a sports perspective, this situation is not uncommon. This example might have been flagged due to e.g., corner-kicks being underrepresented or the gap in time and location between the corner-kick and the action prior to it.



(a) Annotation errors

(b) Rare successful sequence

(c) Rare successful shot

(d) Flagged normal example

Fig. 1. Four examples of flagged sequences. (a) Sequence with incorrectly annotated dribbles (FAWSL). (b) Rare successful sequence resulting in a goal (EURO). (c) Rare successful shot from the halfway line (EURO). (d) Corner-kick that is flagged as abnormal (FAWSL).

5 Conclusion

By inspecting the output configuration of tree-based action-value models, we can detect examples for which the model's predictions may not be trustworthy. On the one hand, we can identify annotation errors, for which the model's predictions are inherently wrong. This could help data providers automatically verify their data before release. On the other hand, we can identify infrequent examples that are dissimilar to the training set. The model’s value should be treated with caution in these situations.

Acknowledgements. This work was supported by the Research Foundation – Flanders under EOS No. 30992574. We thank StatsBomb for providing the data and Jan Van Haaren for providing the videos.

References

1. Baboota, R. and Kaur, H.: Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting, 35(2):741-755 (2019).
2. Shi, Z., Moorthy, S., and Zimmermann, A.: Predicting NCAAB match outcomes using ML techniques – some results and lessons learned. In Workshop on Machine Learning and Data Mining for Sports Analytics (2013).
3. Jaspers, A., Op De Beéck, T., Brink, M. S., Frencken, W. G., Staes, F., Davis, J., and Helsen, W. F.: Relationships between the external and internal training load in professional soccer: what can we learn from machine learning? International Journal of Sports Physiology and Performance, 13(5):625-630 (2018).
4. Op De Beéck, T., Meert, W., Schütte, K., Vanwanseele, B., and Davis, J.: Fatigue prediction in outdoor runners via machine learning and sensor fusion. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 606-615 (2018).
5. Decroos, T., Bransen, L., Van Haaren, J. and Davis, J.: Actions speak louder than goals: valuing player actions in soccer. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 1851-1861 (2019).
6. Robberechts, P. and Davis, J.: How data availability affects the ability to learn good xG models. In Workshop on Machine Learning and Data Mining for Sports Analytics, pages 17-27, Springer (2020).
7. StatsBomb: Introducing on-ball value (OBV). https://statsbomb.com/articles/soccer/introducing-on-ball-value-obv/ (2021).
8. Van Haaren, J.: Why would I trust your numbers? On the explainability of expected values in soccer. In Workshop on AI for Sports Analytics (AISA), pages 1-8 (2021).
9. Van Haaren, J.: https://twitter.com/JanVanHaaren/status/1511003282868781063
10. Hooker, G.: Diagnosing extrapolation: Tree-based density estimation. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 569-574 (2004).
11. Devos, L., Meert, W., and Davis, J.: Adversarial example detection in deployed tree ensembles, arXiv (2022).