CSCI555-Data Mining and Machine Learning

**Date: 9th March 2020**

# A Review Paper on
# Cracking the Black Box: Distilling Deep Sports Analytics

Shubham Deshmukh

Supervised by: Dr. Othman Soufan
Department of Computer Science
**ST. FRANCIS XAVIER UNIVERSITY**

## WHY THIS PAPER?

The paper "Cracking the Black Box: Distilling Deep Sports Analytics" explains the methodologies which can be used to predict outcomes and decisions in the field of sports. I choose this paper to review because of its emphasis on using machine learning in sports for deep and describe analysis, Sports has always been fast-paced and unpredictable, with a good room of machine learning models to predict outcomes and situations. Furthermore, this paper shows a brief comparison of Trees and Neural Networks while suggesting a new solution that can be a great learning curve and can be applied to several machine learning problems.

## PAPER SUMMARY:

The paper "Cracking the Black Box: Distilling Deep Sports Analytics" explains and gives a possible solution to sports analytics while developing a simple machine learning model like a tree that mimics the output of a complex neural model which are used in sports analytics. The paper also explains a trade-off between Transparency and Accuracy while using trees or deep learning while predicting output in sports and ways to eradicate them using Mimic Learning. Mimic learning converts a black box model into a white-box model. To develop a similar performing model, various data augmentation algorithms are applied to the model making it a heuristic and transparent model (Iterative Segmented Regression was recommended). Lastly, the paper also gives a brief comparison between the outputs of a complex neural network and the mimic model developed in the solution. It also discusses important factors like Fidelity, Feature Importance, Rules, and Computational requirement & feasibility.

## POSITIVES & NEGATIVES:

- The paper compares both the black box and white box models very concisely in the real-world usage in sports analytics.

- The paper provides a suitable solution (mimicking model) of problems posed by complex deep learning models to gain both transparency and accuracy.
- The paper signifies the importance of features which can only be achieved by using a tree-based model, the features can then use to obtain real-world rules and having insights into the DRL model.
- With data size in millions, a tree-based model can be affected by overfitting which can hamper the fidelity.

## DETAILED COMMENTS:

Mimic learning has previously been successful in learning shallow black-box models to produce similar outcomes as their counterparts, A linear tree model shows added expressive power and deals with reinforcement learning problems and is being used here instead of regression tree, Several heuristic methods are also experimented and "Iterative Segmented Regression" was found the most efficient as stated in the table. 1, An action-state pair value ($St$ ,$At$ ) is fed the neural network which produces three action-value outputs which are :
1st – Home team getting a goal, 2nd – Away team getting a goal. , 3rd – Match ends as a goal-less draw.

Another important factor derived is Impact, which is calculated at each leaf of the tree by using an Impact function.

$$Impact\ (St, At\ ) = Q(St\ ,At\ ) - Q(St{-}1,At{-}1).$$

The higher the impact value, the more the chances of a goal, Since the impact function also depends on the previous action-state, the impact value of two different goals can be different each time.

The neural network has multiple layers, To make the dataset compatible with the mimic model, each match is divided into episodes such that every episode has its own rewards/penalties. A Q-learning approach has been used to reinforcement learning, where the agent learns iteratively to perform better. SARSA an on-policy algorithm is being used to find a Q-function.

The action in the data depends on timestamps, several layers are generated for action with the same timestamp, and the split is based on a true-false decision, The decision is made at the leaf of the tree.

Data augmentation is very critical to improve the efficiency of a model, A new data augmentation technique is built for this purpose referred to as action replacement, where for any given target action we select any state-action pair and ask the neural network to obtain a soft label, this gives the model to learn extra sets of data to learn from. Also, due to large the data size the trees are grown endlessly and there must a threshold or breakpoint for splitting the trees. Several heuristics for splitting are used for the datasets and "Iterative Segmented Regression" came out to be the fastest, splitting the reduces the y-variance, the splitting criterion is.

$$\sigma\ (\mathrm{s}) - [\ (N\ \mathrm{st}/N\ \mathrm{s}) \cdot \sigma\ (\mathrm{st}\ ) + (N\ \mathrm{sf}\ /N\mathrm{s}) \cdot \sigma\ (\mathrm{sf}\ )]$$

here s is the whole set of data records (xi) on a node, st is the set of data records on a child node for which the split condition is true (xi ≤ ci), sf is the set of data records on another child node for which the split condition is false (xi > ci), and Ns represents the number of data records in set s.

A problem with a growing tree by variance reduction can be overfitting, a solution to this can be to have a proper pruning criterion at different phases this will lead to better fidelity and makes the tree

less complex. The pruning criterion for a tree node v, let Nv be the number of data records assigned to v. Consider the parent v of two leaf nodes v1 and v2. the pruning criterion is E, so if Ev < Ev1 +Ev2, then we prune the two leaf nodes and make v a new leaf node. It goes until Ev >= Ev1 + Ev2 for all leaf node parents v.

## RECOMMENDATION:

I strongly recommend this paper for using machine learning for sports analytics and for showcasing the idea of mimic learning to get fidelity of black-box model and transparency of white-box model.

## REFERENCES:

Xiangyu Sun, Jack Davis, Oliver Schulte, and Guiliang Liu. 2020. Cracking the Black Box: Distilling Deep Sports Analytics. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (*KDD '20*). Association for Computing Machinery, New York, NY, USA, 3154–3162. DOI:**https://doi.org/10.1145/3394486.3403367**