CS 5525 Proposal – European Soccer Predictions

Harsh Patel, Ayaan Kazerouni

Description of the Data: We have chosen a European soccer dataset [1] that contains data about more than 25000 soccer matches and more than 10000 players and their attributes. Covering 11 European countries and their championships, this dataset spans an 8-year period (2008 - 2016). It provides detailed match events such as goal types, possessions and more, for more than 10000 matches. Player and team attributes are sourced from the popular EA Sports FIFA video game series.

The data is provided in a relational database (RDB) of *Matches* (115D), *Players* (7D), *Player-Attributes* (42D), *Teams* (5D), *Team-Attributes* (25D), *Leagues* (3D), and *Countries* (2D). Of these, the entities of interest are Player-Attributes, Team-Attributes, and Matches. We think the other entities will be of use for mostly indexing purposes.

Objective: Our initial goal is to train a model which, given current attributes for two teams, is able to predict the outcome of some future match between them. This prediction includes predicting the winner of the game (if any). Following this, we will also attempt more detailed predictions about a match, such as the number of goals scored. Predictions like these have implications on bookmaking (illegal in the US, reuglated in Europe).

Reasons for Choosing this Dataset: This dataset is interesting because soccer is a widely followed sport and is arguably the only globally played sport. We also think it would be interesting to come up with a model to predict match winners, key player attributes, and league winners. The provider of the dataset mentioned a match-prediction accuracy of 53%. Even though we are unlikely to succeed, we think it would be interesting to try to beat that percentage.

Kaggle provides for iterative improvements of uploaded datasets. The most recent versions were uploaded within 10 days of this writing (versions 7 through 10). The frequently updated nature of this dataset suggests that care has been taken to maintain the high quality of the data.

Proposed Data Preprocessing Steps: Considerable efforts have been put in to collect data from various sources and compile a consolidated dataset and pre-process it by the provider (Kaggle user Hugo Mathien). Thanks to him, we are afforded the ability to issue straightforward SQL queries in our data aggregation steps. This will enable us to reduce dimensionality to easily focus only on areas of interest.

We aim to follow standard data cleaning processes as studied in the class. We will be ignoring Nominal data, such as player ID, team ID, etc., for the purpose of generating the model. We will apply dimensionality reduction techniques to find the most promising team attributes. Again, following the standard practice, we'll come up with attributes that represent more than 90% of

the variance in the data. At this stage we have not yet decided if we will incorporate Principal Component Analysis (PCA) or Singular Vector Decomposition (SVD).

Proposed Data Mining Approaches: For model creation, we are partial to decision trees over rule-based classifiers. With rule-based classifiers, there is a bit more overhead with ordering and prioritising rules that is not present when using decision trees. This overhead comes without an increase in speed of classification. We aim train our model using 70% of the data, and use 10% as a validation set and 20% as the test set. However, this may not be possible. At first glance, the size of the data is sufficient to simply do a 70-30 split for training and testing. However, after aggregating data from many tables, and accounting for severely incomplete rows, the size of the data might reduce signficantly. In this case, we will have to adapt our training and testing techniques.

Depending on the prediction capability of the model, we will move towards developing another model using a naive Bayes approach and compare its results with the former model, if time permits.

Future Work: As proposed, we are limiting our approach to decision trees and naive Bayes. But, given the nature of data, the dependencies between different data attributes cannot be denied. There could be dependencies between team attributes and player attributes. There could be dependencies between attributes of different players as well. A particular player may be very well coordinated with a particular player, but may not be the case with some other player. Hence, the coordination between players could be a deciding factor in a match.

With these dependencies in mind, it would perhaps be better to consider Bayesian networks for the second model.

References

[1] Hugo Mathien. European soccer database. `https://www.kaggle.com/hugomathien/soccer`. Accessed: 2016-10-25.