University of Pisa

DEPARTMENT OF COMPUTER SCIENCE
Master's Degree of Computer Science
(Artificial Intelligence)

Artificial Intelligence Fundamentals

Football-betting Detection System

Project made by:
Andrea Tufo

Professor:
Vincenzo Lomonaco

Academic year 2022/2023

Contents

Chapter 1

Introduction

In this document all the specifications of the project can be found, including not only theorical explainations, but also instances, useful to enucleate the code.
The first chapter is going to introduce how the projects actually works, and why it would be useful for football and more in general for sport. This section is also important to figure out the idea behind the algorithm and how all the development has been organized.
All the main difficulties that I faced during the development and all the most important issues that my algorithm has, are listed in the end of this document, where are explained some possible solutions too, in order to solve them and improve the algorithm.

1.1 Sport-betting

The Sport-betting phenomenon is still nowadays one of the worse side of the sports. It hurts two sports above all: tennis and football. The former because, since there are very few actors involved in the game (for example only two players), it's very easy bribing one of them or both of them and change the flow of events.
The latter because football moves a huge amount of money, thus it's very easy to became millsionaire corrupting one or two match per season.
The *modus operandi* is always the same, "bribe and earn", so "sport criminals" used to corrupt players, who have the role to make the match ends as agreed, then criminals will be able to bet and so collect thier money.

1.2 Modus operandi

In order to choose the right way and so understand what kind of data my algorithm needs to work as expected, I stuidied how criminals operate, trying to find many information as possible on the Internet.

Usually these lawbreakers rig a few matches per division in a season and, furthermore, also few teams are involved, maximum of three or four teams per division. For example in Serie A during the season 2011/2012, only two matches has been faked, for a total earning estimated between 300'000 and 500'000 euros, corrupting players who belong to only two teams. Accordingly, it's quite difficult to detect anomalies observing all the season of a specific team, so it's important to highlight that, especially in this case, technology and human must work together to reach the goal and get rid of false-positive.

Chapter 2

Implementation

For the implementation the main goal was to develop a system that relying and analyzing on matches football data, shows a percentage of "possible unfair match". The basic idea was to retrive some parameters and values from raw filtered data, collect them, and then looking for some anomalies. In the code has been used as case of study the 2012/2013 season of Serie A, so every match that belongs to the regular season. In my case there is one important metric, which is called *"UGI" (Potential Goal Index)* and has the role to mesure how a team has been dangerous and offensive during a match, so this value is computed for each football team.

2.1 Dataset

During the first fase, after the preparatoty study of the phenomenon, was to select a reliable dataset from the Internet, so in the project my choise was to pick up a dataset found on kaggle, which contains six seasons matches' data of five championships. Of course these raw datas were no ready to be used for my program, thus, filtering them and select only the needed ones was the first operation that I implemented. The dataset is made up two csv files, one "ginf.csv" that contains all the matches (with home and away team name, goals, stadium, season, league), while the second one called "events.csv", contains all the events per match and so fouls, shots, ball possesion losses, corners, penalties and more. These two files are connected like in DBs, every match in "ginf.csv" has an unique id, that identifies it in the other file.
All the first effort was focused on filtering all the data of the 2012/2013 italian season, collecting 380 matches, with all statistics and events.

2.1.1 Events file

The events file has three main columns: the event_type, which throught an index give us the information on what type of event has happened (for example 1 for shot), than there is the event_team that contains the name of the team that generates that specific event and finally the location column which indicates the position of the events throught a sort of field mapping.
Of course there are many more attributes that are very important like the final outcome of the shot (goal, post hit, blocked not in target), or the description of the event but we will focuse only on this three parameters because are the most used in the code.

2.2 Filtering

The filtering phase has been very hard to implement because only during coding the algorithm logic I figured out step by step what kind of data my program needed. So the main operation could be divided into two parts: the first one in which I filter the useless matches, selecting only the matches on which I was intrested to work on, while during the second part my goal was to select only the correct datas from the events file. The data selected per single match were: the goals scored by home and away team, the victory, draw and loose probability, the number of shots, the locations of evry single shot excluding the shots that had as outcome "goal scored".
During the first run of the algorithm a "filtered_dataset.csv" file is generated, it contains all the data filtered in order to avoid to fetch and to filter more times the data from the biggest file, because it takes too much time to do it.

2.2.1 Potential Goal Index

Pratically in my program I worked on only one json object, which contains all the data filtered and makes more easy get all the metrics from every single match. The biggest challenge for me was to find out some parameter that could highlight how much offensive and dangerous a team is during a match in average. And my idea was to get analize every match looking at the shots made by a team and them locations. So for all matches I group the two arrays of shots locations (one for home team and one for away team) into three group, each of this group has a weight that refers to the possibility to score, thus, fisrt group has weight one, so very low dangerous shot, the second one has weight two, and the third one has three as weight, and so high chances

to score.

Then I caluclated the cardinality of these three sets and after this I computed the *UGI* summing the cardinalities timed for the weight of the group on which they belong to.

Chapter 3

Final results and HMM application

The reason why I choose Hidden Model Markov as probability algorithm is due to how the dataset is made up. During the entire work I have never tought about what kind of algorithm I was going to use. I simply tried to elaborate datas thinking only on what can be usefull and what would not be used in the program. Finally when all the values, metrics and parametrics has been obtained from the files, I started to think on what algorithm I could use.
I had two main parameters parameters per match: goal scored by the two team, the home and away team UGIs and the.