DATA MINING IN SPORTS BETTING

Cristian Georgescu
cristian.georgescu@upcmail.ro
„Dunarea de Jos“ University of Galati,
Faculty of Economics and Business Administration,

In this paper, we have made a brief analysis on how to make decisions in betting on European football with the help of data mining techniques. Whether you refer to betting a few days in advance of the sporting event or live betting, both options have been taken into consideration. By using a clustering algorithm for analyzing both the database containing events from football matches and the odds given by bookmakers, we have obtained graphs indicating the probabilities associated with analyzed events. Given the purely informative aspect of the current paper, we have only analyzed the number of corners from a match.

Keywords: betting, European football, data mining, clustering

1. Introduction

It is well known that in these days sports betting are approached using the combined power of databases and mathematical analysis tools for establishing both betting odds and options for betting. This paper is an argument for the advantages of using data mining algorithms over the usual statistical tools currently used in this domain. For exemplification of using data mining analysis tools, we have used a clustering algorithm for predicting the number of corners both before the match, and live, during the match, conditioned by other events already produced.

As is the case for business, the decision for betting on a specific sporting event can sometimes involve important sums of money, so there has to be a decision making process based on objective components. Oftentimes, professional punters base a lot of their bets on experience, intuition, knowledge of the sport and a few formulas they have deduced along the way.

In (Schumaker, Solieman, & Chen, 2010) five analysis levels for making this sort of decisions are synthesized: no substantiation, experts who know the domain and make predictions on instinct, experts basing their decisions on historical data, using statistical data in decision making and using data mining algorithms for supporting decision making. In this paper we consider the highest level of analysis: data mining.

The basic principle of winning from betting uses a formula of the following type (Olesen, 2008):

\[ V_i = \text{odd}_i \times \text{prob}_i \]

where odd\(_i\) is the bookmaker’s odd for event i and prob\(_i\) is the probability that event i will happen. If \(V_i > 1\) then there are winning chances, but if \(V_i < 1\) then no betting goes on.

But the most challenging and widespread betting setup is online betting on European football matches, where live stats are updated on the website along with odds for future events from the game. In these cases, in a certain minute of the game, time left in play is estimated based on the referee’s inclination to give added time. Time left in play, teams’ style of play, events produced and other elements allow the punter to calculate certain, personal probabilities with which future events may happen. If these probabilities are superior to the odds given by the website, then bets are placed. In this situation, when the punter places more bets on different events, the above-mentioned formula is added up for all bets.

\[ \Sigma V_i = \Sigma \text{odd}_i \times \text{prob}_i \]

Inevitably, there will be certain situations when, for a specific i, although \(V_i\) was greater than 1, it became smaller than 1 as events happened along, which forces the punter to introduce compensation bets, the purpose being to maximize the amount of money won from the bets.

In online betting, the punter has limited time available to place a bet whenever they find better probabilities, because odds are quickly updated according to events from the live game. Therefore, a race
between who calculates events' probabilities quicker is on, results' performances relying on the quality of current and historical input data, but also on the algorithms used.

In the analysis process presented in this paper we have used a database containing almost 36,000 games, totaling over 2.8 million events occurred in these games (corners, fouls, yellow/red cards, goals, etc.)

2. Exploring through Data Mining with clustering algorithms

In approaching the analysis of sports betting domain from this article, we have used a clustering algorithm form Microsoft SQL Server Standard Edition 2008 R2. This clustering algorithm is from the unsupervised classification category of algorithms. As opposed to supervised classification where elements subject to classification are labeled (pre-classified), clustering represents the process of distributing elements to be classified into clusters in an unlabeled manner (Jain, Murty, & Flynn, 1999).

In SQL Server, there are two options for the clustering algorithm: K-means and Expectation-Maximization (EM).

The name of K-means algorithm comes from the division into k clusters of Cj where j = 1,...,k and their mean is c_j. The value of parameter k is set at the start of running the algorithm and the value c_j is the calculated weight centre of points from C_j. This weight centre is named centroid (McCaffrey, 2013) and it defines the most representative tuple from the set, chosen by calculating the distance to the other tuples. Because it distributes elements into disjoint clusters, the K-means option is also called hard clustering (Microsoft, 2013).

The EM option does not calculate geometrical distances, but uses a probabilistic method in distributing elements into clusters (Dempstern, Laird, & Rdin, 1977). It is an iterative algorithm, where each iteration has a preparation step in which missing data is estimated (expectation step; E-step) based on observed data and parameters of the model, while in the second step, using a probability function, elements are distributed into clusters, assuming data estimated in the first step is known. Distribution is done by maximizing the probability function (maximization step; M-step). These two steps name the algorithm Expectation-Maximization (Jiawei, Kamber, & Pei, 2012).

3. Betting on total number of corners

The case used for exemplification is the total number of corners in a game. The typical bet in this case is: "under/over 11.5". As the game goes along, according to occurred events, other options may become available, like “under/over 9.5”.

A base for analyzing a substantiated answer can be found in the graph from Figure 1, obtained by using the prediction function of the clustering algorithm.

At a first look, it seems a predictable result, being a normal distribution common to many statistical phenomena and obtainable by calculating simple means of the total number of corners. In reality, resulted probabilities are conditioned by odds given by bookmakers for the final result of the match. Specifically, it is about the away early odd, meaning the chances of winning the game given to the visiting team with two or three days ahead of the match. Similarly, you can obtain this type of graphs according to the home team’s odd or according to the odd for a draw game. You can also replace the odds with those given just before the game starts.

![Figure 1: Probability of total number of corners in a match, according to away early odds](image)

The betting offer, calculated by bookmakers, for total number of corners from a game is made in such a way as to mislead punters. Ultimately, this total number is the sum of corners taken by the visiting
team and by the home team. A more careful analysis (Figure 2) shows us that in reality there are two separate events that rely on different factors. The arrows in Figure 2 that represent dependencies are characterized by different probabilities that offer information on the strength of these dependencies. It is not surprising that the league in which the match is played determines both the total number of corners and, separately, the number of corners taken by the home and away teams.

But dependencies between the away team’s number of corners and the home team’s name and also between the home team’s number of corners and the away team’s name are more complicated to anticipate. In other words, the total number of corners depends on the opposition, a fact confirmed when taking a closer look at the game because, from this point of view, the manner in which a team defends is more important, than the manner in which it attacks.

Figure 2: Dependencies between the number of corners and the name of the league or teams

Having a database with historical data of events from each match at our disposal, a database that is accordingly organized for the algorithms used, we pose the question of which interrogations might we formulate so as to provide information regarding the most probable events from a match.

From the multitude of possible interrogations, which depend on the ability of each punter, we have only chosen one: “How is the total number of corners influenced within a game and by what?” Using this question as the starting point, there are a lot of situational variances: “How is the total number of corners influenced when the home team is the favorite?” which means the home team had odds to win between 1.00 and 1.40 before the game or “How is the total number of corners influenced when the away team is favorite?” which means the away team had odds to win between 1.00 and 1.40 before the game. As we have already shown, the total number of corners must be obtained by adding up the separate analyses of corners taken by the home side and those of the away side.

Coming back to the online aspect of betting, the most important situations are those that take place during the game because punters must adapt in real time to the new context. Such an example would be interrogating “How is the total number of corners influenced when the home team is the favorite and the visiting team scores in the first 30 minutes of the game?” Using the same algorithm as for Figure 1, we obtain the graph from Figure 3, which in this situation has significant deviations from the normal distribution obtained in the former case.

Figure 3: Probability of total number of corners when home team is the favorite and away team scores in the first 30 minutes
Variations of interrogations for the number of corners can be done at any moment of the game, conditioned also by other events that have already taken place or by a set of events. When providing the answer, you should also take into account all corners that have already been given for either the hosts or the visiting team. This can be done either simplistically, by subtracting one from the other, or more elaborate, by considering corners already executed as events from the match and calculating probabilities conditioned by these events.

4. Conclusions

As we have previously mentioned, when it comes to online betting, a crucial factor is the speed with which the punter calculates, at a certain moment of the game, probabilities associated with future events.

Literature from this domain states that the total data mining time for an economic problem in sales with 10 million transactions is more than a few tens of minutes, when data is organized in a standard relational manner, while it only takes a few seconds should the data be organized in a hypercube. In the case of events from football matches happening either in professional leagues or international competitions over a few years, you can easily gather similar volumes or even higher volumes of data.

Thus, the issue of organizing data becomes as important as having the right data and algorithms. As is the case for transactional systems, organizing data in a hypercube is a must. The high-performance hypercube data organization is based on the most frequent and probable interrogations in order to obtain the quickest possible response time. This performance is optional in the case of standard betting, which has no time restrictions, but is crucial for online betting.

Bibliography