Applying Data Mining Techniques to Football Data from European Championships

Sérgio Nunes[1] and Marco Sousa[2]

[1] Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto, Portugal
sergio.nunes@fe.up.pt
[2] zerozero.pt
http://www.zerozero.pt
msousa@zerozero.pt

Abstract. Data Mining is the process of finding new, potentially useful and non trivial knowledge from data. Football is a popular game worldwide and a rich source of data. Gathering only part of this data we are able to collect hundreds of cases. In this paper we describe an exploratory work where we use Data Association Rules, Classification and Visualization techniques to find patterns in datasets from several European championships. For each one of these techniques, different hypotheses were stated. For Association Rules and Visualization, our hypothesis was that we would be able to find non trivial knowledge and confirm several known patterns. For Classification, our hypothesis was that we would be able to classify matches according to their results based on the available history. Our findings didn't confirm our hypotheses to their full extent as expected. Our exploratory work confirmed several well known patterns in football and highlighted borderline cases. Among the several techniques used, visualization produced the best results.

1 Introduction

1.1 Context and Motivation

Data Mining (DM) is commonly viewed as a specific phase in the Knowledge Discovery in Databases (KDD) process. Currently, *Data Mining* is an overloaded term used to mean several concepts. We consider DM to be the application of machine learning techniques to extract implicit, previously unknown, and potentially useful information from data [12]. Nevertheless, during this paper, we will sometimes use this term to refer to the whole process of KDD. The exponential increase in the amount of data that exists stored in electronic databases has fostered the growth of this field. A simple search for "data mining" in any popular web search engine will return several millions of hits [1].

[1] In December 2005, a search in Google Search returns more than 17.900.000 hits.

Football is a very popular game worldwide, it was invented in England in the XIX century and is now played regularly by more than 240 million people according to Fédération Internationale de Football Association (FIFA) [8]. Football is also known as soccer, or association football, in some countries, namely in the USA.

The motivation for this project arouse from an opportunity to work with a large database of football data. This data was provided by zerozero.pt [4], an independently maintained website that gathers and presents data from several football championships worldwide. The data granularity varies significantly among championships. Two main datasets were used. The first dataset includes the 2004/05 edition of the Portuguese championship and was chosen because it is the one with the highest level of detail and the lowest levels of missing values and erroneous data. The second dataset includes all matches played in six European countries, including Portugal, for the last 50 years [2]. Although rich in the total number of cases, this dataset has very few attributes available.

In this paper we present an exploratory work where we apply several DM techniques to the chosen datasets in search of existing patterns. We expect to find patterns that relate the events in a non trivial fashion. If these patterns are found, they can provide valuable insight to the people involved directly or indirectly in the match. An example of application would be the development of a decision support system to be used during the match. Another example application would be the use of this information to aid in the selection of referee or locations for each match.

We are not aware of any published work where these specific DM techniques are applied to football data to discover or confirm existing patterns. Research found in this area is mainly related to robot soccer and autonomous agents [10]. In this case, data mining modules were developed to provide adaptive agent behavior in dynamically changing environments using automata data. Considering the use of DM in other sports besided football, the work published by Bhandari et al. in 1997 [6] describes Advanced Scout, a PC-based data mining application used by the NBA coaching staffs to discover interesting patterns in basketball game data.

1.2 Paper Structure

The Cross-Industry Standard Process for Data Mining (CRIPS-DM) [7], an European Community developed standard framework for data mining tasks, identifies six generic phases in the life cycle of a data mining project. In this work, these phases are used to structure the paper. The first phase, called **Business Understanding**, focuses on understanding the project objectives and requirements from a business perspective and setting a preliminary plan to achieve the objectives. This has been covered in Section 1.

In Section 2, the next two phases of the CRISP-DM process are covered, **Data Understanding** and **Data Preparation**. The Data Understanding phase starts

[2] For some countries we have all the matches played since the XIX century.

with the initial data collection and proceeds with activities in order to get familiar with the data. Also included in this phase are activities related to the analysis of quality problems in the data. The Data Preparation phase covers all activities to construct the final dataset from the initial raw data. Also in this section, initial obvious results are presented.

In Section 3, the **Modeling** phase of the CRISP-DM process is covered, several modeling techniques are applied and tuned for best performance. We use Data Association Rules, Classification and Visualization to mine the datasets.

Finally, the **Evaluation Phase** is covered in Section 4. The results from the previous section are organized, presented and discussed. In this section our work is viewed in the light of our initial hypotheses.

In the CRISP-DM framework, one last phase of deployment is identified. This phase wasn't included in our work since it wasn't one of our goals.

2 Datasets

2.1 Data Preparation and Exploration

The raw data was collected from zerozero.pt's [4] main database. The data is stored in a relational database management system and was exported to flat CSV files using PHP scripts and SQL. The initial exploration and preparation of the CSV files was done using R [2], an open-source language and environment for statistical computing and graphics.

The two datasets used are described in the following sections. Initial data explorations are also described.

2.2 Portuguese Championship 2004/05 Events

All existing events from the 2004/05 edition of the Portuguese football championship were exported. This edition of the Portuguese championship included 18 teams that performed a total of 306 matches, there were 711 goals scored and a total of 1.771 cards shown by the referees.

The exported data includes information about the players in each match, substitutions made during the match, the time and location of the match and information about the teams and players when the match happened. For instance, for each team, there is information available about the number of points, goals scored and goals conceded since the beginning of the championship. On the other hand, for each player and besides demographic data, there is information about the number of goals scored and cards received since the beginning of the championship.

The final dataset has more than 17.000 cases, each one with more than 50 features. Each case represents an event (see Table 2). For each event, the available features are summarized and explained in Table 1.

Football occurrences stored in the original database were analyzed and normalized to fit a standard representation. In this standard representation, each

Table 1. Features available in the Portuguese Championship dataset.

Group	Related Features
Event	Related to each event: type, minute and half within the match.
Match	Related to the match being played: date, start time, score, TV channel transmitting, referee, number of spectators and total overtime granted.
Teams	Related to each team involved in the match: name, coach and current position, number of points, victories, defeats and draws in the championship.
Location	Related to the place where the match takes place: stadium and city.
Player	Related to the player involved the event: name, age, playing position, nationality, birth country, weight and height.

event has only one player associated. Hence, all occurrences were split in multiple simpler events. For example, a substitution corresponds to 2 events - one associated with the player leaving, another associated with the player entering. Another example is the initial line up of the teams, that correspond to 36 events - 11 starter events and 7 substitute events from each team. In the end, 9 types of normalized events were identified and characterized. These types are depicted in Table 2.

The final dataset has very few errors or missing values. This was one of the factors considered to choose this dataset. In fact, the 2004/05 edition of the Portuguese championship is the most complete one in zerozero's database. Existing errors, missing values and outliers were easily detected using simple statistical tools, namely boxplots. These records were deleted, no attempt was made to fill in or correct the data.

An initial exploration of the data was performed using statistical tools. A density chart for event types was plotted (see Figure 1). It is interesting to note that:

- Substitutions only start to occur at the end of the first half, being rare at the beginning of the match.
- There is a strong peak of substitutions near the minute 45. This corresponds to the substitutions performed at half time.
- The number of cards shown increases during the match with peaks at the end. Red cards and second yellow cards have a very high peak near the end of the match.
- During the first half of the match, the number of double yellow cards is very low and is surpassed by the number of red cards.
- The number of goals doesn't exhibits peaks but increases in the end of the match.

Table 2. Event types in the Portuguese Championship dataset.

Event Type	Description
Starter	Represents a starter player included in the initial lineup. For each match there are 22 events of this type, 11 for each team, occurring in the minute 0 of the match.
Substitute	Represents a substitute player for the match. For each match there are 14 events of this type, occurring in the minute 0 of the match.
In	Represents the exiting of a player during a substitution.
Out	Represents the entering of a player during a substitution.
Yellow	Represents the showing of an yellow card to a player.
Second Yellow	Represents the showing of the second yellow card to a player.
Red	Represents the showing of a direct red card to a player.
Goal	Represents the scoring of a standard goal.
Penalty	Represents the scoring of a penalty.
AutoGoal	Represents the scoring of an auto goal.

Although a football match starts at minute 0 and ends near minute 90, in Figure 1 the various lines begin before and end after these values. This is a result of the smoothing performed by the density function available in R.

2.3 European Matches

The second dataset contained information about the championships and matches from several European countries. The countries included were: Portugal (15.382 matches since 1934), England (43.730 since 1888), Spain (19.846 since 1930), Italy (17.680 since 1946), France (22.702 since 1933) and Germany (13.406 since 1963). Although a large number of cases (matches) were collected (132.749 in total), few features were available for all matches for all countries. The features included in this dataset are shown in Table 3

In Figure 2, the three major teams in Portugal were plotted by year and by final position. Each team was drawn with a different shade of gray. It is evident the predominance of these three teams in the history of the Portuguese championship. A more detailed analysis of this figure reveals that:

- FC Porto has the most irregular path. Prior to the 80s several fluctuations in the final position achieved are evident. While Benfica has the most overall consistency.
- The 50s were dominated by Sporting, the 60s, 70s and part of the 80s were dominated by Benfica and, since the middle of the 80s, FC Porto has won

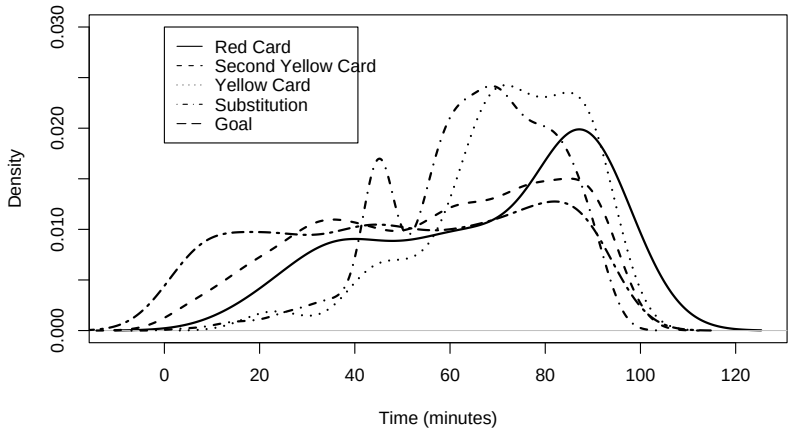

Fig. 1. Density plots for event types.

Table 3. Features available in the European Championships dataset.

Feature
Visited and Visiting team's name.
For each match, the number of goals scored, the number of goals suffered and the winner.
Country's name, year and decade of the match.
For each team, the number of goals scored and suffered for each specific championship (total, in and out).
For each team, the number of points, victories, draws and defeats for each specific championship (total, in and out).

most of the championships. A density plot for each first place for each team clearly reveals this pattern.
– For each team, exceptional bad seasons are evident - FC Porto (40s, 1969) and Benfica (2000).
– The two championships won by none of these three teams are easily spotted, 2000 (Boavista) and 1945 (Belenenses).

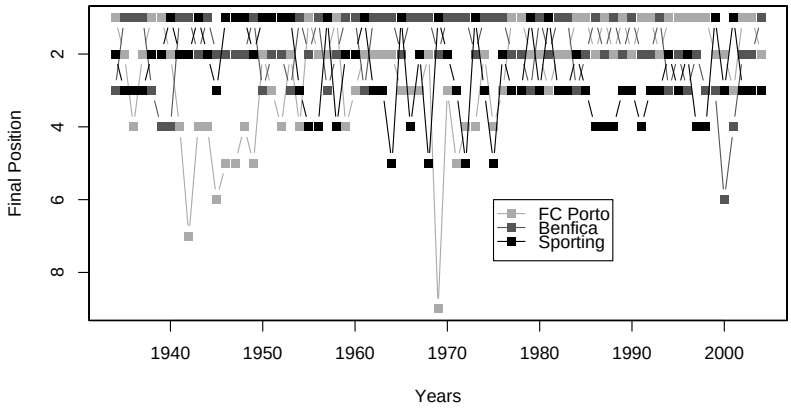

Fig. 2. Final positions for the three major teams in Portugal (1934-2004).

Among countries, density plots for the matches along the years reveal interesting patterns. In Figure 3, density plots for England, France and Portugal are shown. Before 1920 and after 1940 the two World Wars are evident in the plots for England and France. For Portugal, the increase in the number of matches is visible.

3 Modeling

3.1 Association Rules

Mining for association rules is a DM technique that enables the finding of frequent patterns, associations, correlations or casual structures among sets of items. This task was performed using two different open-source software tools, Weka [3] and AlphaMiner [1]. Due to the low number of attributes in the European championships dataset, only the Portuguese dataset was used. In this case, after the discretization of numerical variables, a total of 40 nominal attributes in 16.900 cases were available.

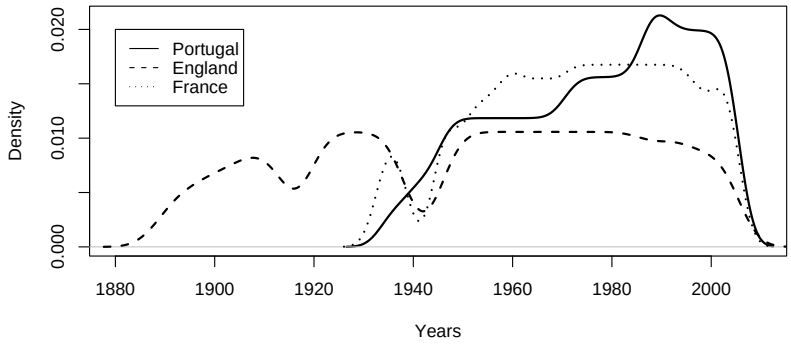

Fig. 3. Total number of matches density.

We used the Apriori algorithm [5] to search for association rules. Three types of metrics were used with different minimum values: Confidence (75%), Lift (1.5) and Leverage (0.1). For each one of these metrics, a minimum support of 25% was set and a maximum of 100 rules were produced.

Having a reasonable number of attributes and a high number of cases yielded high expectations towards the finding of patterns. Nevertheless, after exhaustive exploration, no interesting or unexpected rules were found. Only trivial rules were identified, for example: *"Matches that start between 15:30 and 16:30 are on Sundays"* (84% conf.) or *"Matches that are not transmitted on TV are on Sundays"* (80% conf.).

3.2 Classification

Classification is a DM technique for mapping objects into predefined classes. Classification was performed using Weka's implementation of the C4.5 algorithm [9], named J48. This technique was used only with the second dataset. In this case it is possible to set interesting, realistic and useful goals. Despite having many more attributes, the first dataset is less interesting as a classification problem. In this case, simple tests have showed that, for example, predicting the end result of a match is quite trivial since we have all the events for that match.

With the second dataset, including match results from several European countries, we set the goal of classifying each match according to the final result. Three match results are possible for the visited team: victory, defeat or draw. A very small set of attributes was used, namely the name of both teams and the year of the championship. The dataset was also split by country and several runs of the classification algorithm were performed with different values for the confidence factor (C) and the minimum instances per leaf (M). The values used

were: C (0.05, 0.1, 0.5, 1, 10) and M (1, 5, 10, 20, 50). Each model was tested using a training set (70%) and a test set (30%).

For Portugal, the best model (C=0,05 and M=50) was able to correctly classify 59,81% of the test set instances. This score was obtained with two simple rules:

- When the visiting team is "FC Porto", "Benfica" or "Sporting" the result is **defeat**.
- In every other case the result is **victory**.

In this model, no matches were classified as "draw". With a trivial classifier, based on the frequency of each result in the Portuguese Championship (victory 54%, draw 23%, defeat 22%), we have a success rate of 54% classifying every match as "victory". Thus, we can state that our classifier only slightly improves this result, being able to correctly classify 5% more cases.

Nevertheless, only for Portugal we were able to surpass the results achieved by the trivial classifier. In each of the remaining five countries, the best rules simply classified every match as "victory". Thus achieving a success rate equal to the one accomplished with the simple statistical classifier. This can be explained by the predominance of only three teams in the Portuguese Championship. In all other countries there is a greater balance among the various teams, making classification based on a small set of attributes a harder task.

3.3 Visualization

Visualization techniques make use of graphics to produce multiple observations of the data. Of the methods used in this work, this is the most exploratory since no rules are defined on how to conduct research. Visualization is mainly developed for human observation and allows multiple insights into the same data. Visualization can be used to simply view outputs from the application of other techniques or to explore the initial input. In this work, several plots were drawn using R with an exploratory mindset. In this section we show and comment those that are most revealing or unexpected.

Although several experiments were made with the first dataset, visual results were below our expectations. Hence, only explorations with the second dataset are presented. In Figure 4, first places among countries are plotted. Each line represents one country and each year is depicted in the X axis. For each team that won the championship, a different color was used. Different shades of gray were used since they provide an easily comprehended scale to human observation [11]. It is important to note that we only have all the matches, from the start of each championship, for Portugal, England and France. Nevertheless, the following observations are possible:

- Portugal has a very low diversity in the number of teams that won the championships.

– England has the highest diversity on the teams that won the championship. The 50s mark a clear separation on the teams that commonly won the championship.
– Interruptions, mainly due to the World Wars, are easily spotted among championships.

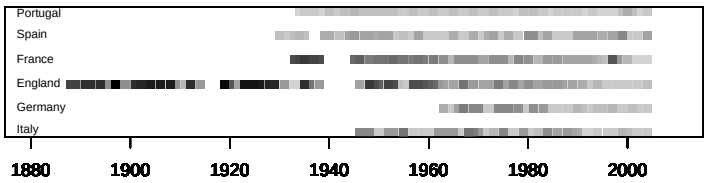

Fig. 4. First placed teams in European championships.

An alternative visual display of the three major teams in the Portuguese championship was produced. In Figure 5 teams are plotted by year and by total points achieved, instead of their final position (as in Figure 2). Although the final positions aren't so clear, more information is available in this second graphic. We are able to see the evolution in the total number of points along the years, reflecting the evolution in the number of teams. Also visible is the increase in the mid 90s, as a consequence from the changing of the rules (victories worth 3 points instead of 2). Excellent seasons are easily spotted, namely Benfica's (1971, 1972) and FC Porto (1995, 1996). Also interesting to note are the bad overall seasons as compared to neighbor championships. For example, in 2004 the winning team achieved fewer points than the third team in several of the previous years.

We've also performed a visual analysis of the evolution of match results for each year in each country. Each match result was plotted in a 2D graphic with the X axis being the goals received and the Y axis the goals scored. These results were then grouped by year and the year's centroid was calculated. In Figure 6 and 7 these centroids are plotted for Portugal and England. Different shades of gray were used for each year, so that the time dimension was visible in the figures. Although similar in recent years, these figures show that match results in England have fewer variations. In Portugal, significant differences between the older matches and the more recent matches are impressive. These analyses were also performed for the other countries and we concluded that France, Germany and Italy exhibit a pattern similar to England's, while in Spain the pattern is more similar to Portugal's.

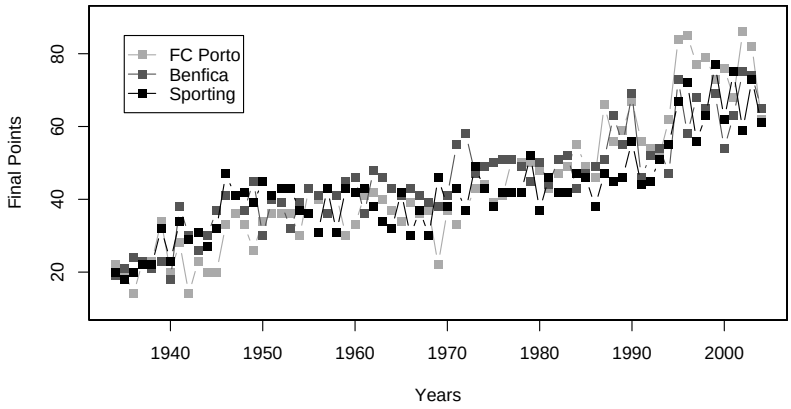

Fig. 5. Total points by championship for the three major teams in Portugal (1934-2004).

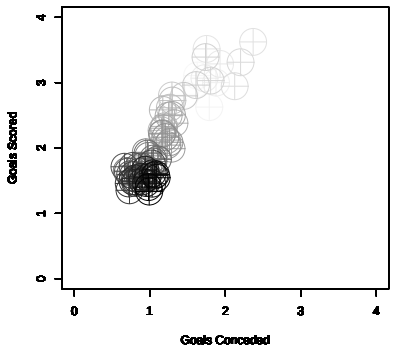

Fig. 6. Centroids for match results in Portugal (1934-2004).

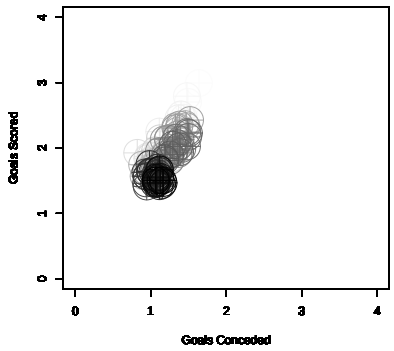

Fig. 7. Centroids for match results in England (1988-2004).

This type of centroid plots were also used to analyze data within each country. Two different analyses are shown for the Portuguese championship. In the first example (Figures 8 and 9), each plot represents the team's match result according to four dimensions: time, goals scored, goals conceded and place. Time is represented using different shades of gray, lighter colors portrait older matches. Goals scored and goals conceded are depicted in the plot's axis. Finally, the place of the match is distinguished using different symbols for each centroid, matches at home are plotted using a circle while matches away are plotted with a square. These plots were produced for every team in the Portuguese championship. Benfica and Boavista were chosen because their plots reveal contrasting evolutions in each team's match results. While Benfica had a greater change in home matches, Boavista had an even greater change in away matches.

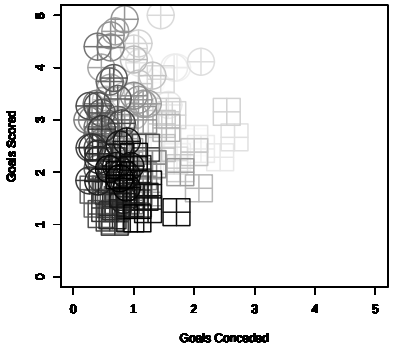

Fig. 8. Centroids for Benfica matches in the Portuguese Championship (1934-2004).

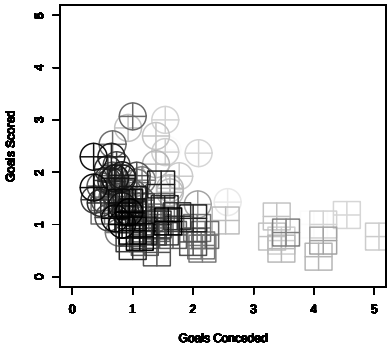

Fig. 9. Centroids for Boavista matches in the Portuguese Championship (1934-2004).

Finally, a similar type of graph was used to compare two teams. In Figure 10 and 11 two examples are shown. For each two teams, the most common match results are shown using different sizes for each point. It is important to note that these plots represent only the matches of Team A versus Team B, not Team B versus Team A. In the examples shown, two different patterns are visible. As expected, in matches against Belenenses, FC Porto concedes fewer goals and the results are concentrated in the "victory side" of the plot. With Benfica, while victories still dominate, draws are more frequent and the amplitude of goals scored is much lower.

4 Conclusions

Our initial expectations were that we would be able to find non trivial knowledge from the available datasets. After several explorations only existing strong sus-

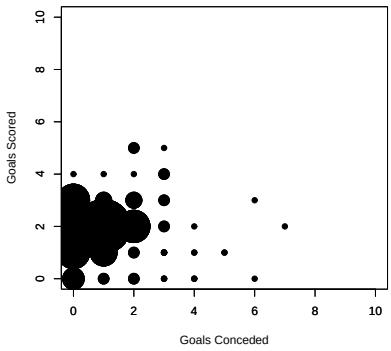

Fig. 10. FC Porto versus Benfica in the Portuguese Championship (1934-2004).

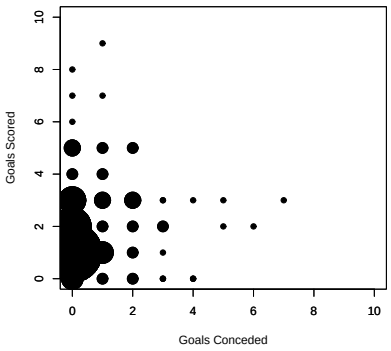

Fig. 11. FC Porto versus Belenenses in the Portuguese Championship (1934-2004).

picions were confirmed. Although we were able to extract knowledge from these datasets, no important and unexpected result was revealed, thus our initial hypothesis was partially refuted. It is important to refer that our hypothesis was partially refused for these two datasets where, although a significant amount of cases is available, the number of attributes is limited. We believe this is the main reason for the bad results obtained with Association Rules and Classification.

With Classification we were able to produce a model for the Portuguese championship that returns better results than a pure probabilistic classifier. This can be explained by the high predominance of three teams that exists in this championship.

The good results obtained with visualization were unexpected. We believe that the high number of numerical attributes and the existing knowledge of the domain greatly justifies this success. Most of the graphics produced emerged as a way to see patterns that were already known in advance. While the other two techniques search for patterns with few inputs from domain experts, with visualization human intervention is necessary during the decision process.

Data preparation is a very time consuming step in the KDD process. Gathering and preparing data to be used with the different algorithms occupied a significant part of the whole process. More than two thirds of our work was invested in data preparation. The word *mining* clearly reflects the nature of the whole KDD process. A lot of time is spent searching for patterns, adjusting parameters in the algorithms and drawing graphics, to find out that only a minimum part of this work is useful in the end. The results obtained are directly related to the time invested in the work.

Several tools were used to perform the data mining tasks. AlphaMiner was found to be very well designed for a knowledge discovery work. Tasks are graphically shown and the steps are evident, useful for the kind of work developed while following an exploratory path. Although being graphically intuitive, this

tool offers less KD methods than Weka and can't cope with large volumes of data as well as Weka. R is an excellent statistical software tool, it is able to perform calculations on large datasets and provides a large repository of packages with extra features.

As future work, we suggest additional exploration of visualization techniques and, if possible, the gathering of more attributes to allow the use of other data mining techniques with improved success. Due to the characteristics of our datasets, we also suggest the use sequential pattern analysis algorithms for finding association rules.

References

1. AlphaMiner. Available from: `http://www.eti.hku.hk/alphaminer/` [cited 2005-11-28].
2. The R Project for Statistical Computing. Available from: `http://www.r-project.org` [cited 2005-11-28].
3. Weka 3 - Data Mining with Open Source Machine Learning Software in Java. Available from: `http://www.cs.waikato.ac.nz/ml/weka/` [cited 2005-11-28].
4. zerozero.pt :: Porque todos os jogos começam assim... Available from: `http://www.zerozero.pt` [cited 2005-11-28].
5. Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules, 12–15 1994.
6. Inderpal S. Bhandari, Edward Colet, Jennifer Parker, Zachary Pines, Rajiv Pratap, and Krishnakumar Ramanujam. Advanced scout: Data mining and knowledge discovery in NBA data, 1997.
7. The CRISP-DM consortium. CRISP-DM 1.0 - Step-by-step data mining guide, 2000. Available from: `http://www.crisp-dm.org/CRISPWP-0800.pdf` [cited 2005-11-28].
8. FIFA. FIFA Survey: approximately 250 million footballers worldwide, 2000. Available from: `http://www.fifa.com/fifa/survey_E.html` [cited 2005-11-28].
9. J. Ross Quinlan. C4.5: programs for machine learning, 1993.
10. Lev Stankevich, Sergey Serebryakov, and Anton Ivanov. Data Mining Techniques for RoboCup Soccer Agents. In *AIS-ADM*, pages 289–301, 2005.
11. Edward R. Tufte. The Display of Quantitative Information, 1983.
12. Ian H. Witten and Eibe Frank. Data Mining: pratical machine leaning tools and techniques with Java implementations, 2000.