# Chapter 5

# Ranking Individuals by Group Comparisons

This chapter focuses on predicting success of the individuals using the likelihood ratio that was introduced in chapter 4. This goal is achieved in three steps. Step one, data is grouped into categories. Categories can be position that players play in (e.g. defender, goalie, striker or midfielder) or the genre of the movie (action, comedy, etc.). Step two, two graphical models are learned. One that represents the population association patterns (e.g. defenders as a whole) and the other one represents individual association patterns. In the next step the likelihood ratio metric is computed using these two models. The computed score is used in order to predict the value of the individuals and rank them. An empirical evaluation on soccer and movie data shows a strong correlation between the likelihood ratio and success metrics.

## 5.1 Introduction

Predicting success especially ranking individuals has been a popular topic for statistics/machine learning research in the recent years.

Appearance of professional soccer statistics websites has made it possible to extend statistical studies to sport domain. One of the very interesting problem in this domain is to predict success and present true estimates of individuals abilities. Market value varies for different players in different age and nationality in different periods of time. And it is worth

to study various factors that could have influence on market value of the players. One way to tackle this problem is to manually aggregate important features of individuals over time and then rank players based on their performance in those features. For example, compare players based on the goals or shot efficiency that they have scored. But this criterion may be unfair to most players as not all the players are in the position to shot or score a goal (e.g. goalies or defenders). One may argue that defenders (or goalies) have some features that are a lot stronger in their group compare to other groups. However, knowing what features are more important for which category of individuals and how to treat the unimportant features is a hard task.

Another disadvantage of manual aggregation is that it does not capture the attributes that represent interaction between players. We later on show that by applying a generative model such as Bayesnet to soccer data we can see which interactive features are useful to process. For instance, number of matches that a player plays becomes important when the player shows high dribble efficiency and pass efficiency.

In this chapter we propose a method to rank individuals which is based on a generative model of the data. The generative model represents the population general behaviour. If the model assigns a low likelihood to generating an individual it says something about that individual and shows how different that individuals is from normal populations. A generic Bayes net (BN) structure is learned with data for the entire population. The nodes in the BN represent features for links, of multiple types, and attributes of entities, also of multiple types.

**Approach** A class-model Bayesian network (BN) structure is learned with data for the entire population. The nodes in the BN represent attributes for links, of multiple types, and attributes of objects, also of multiple types. To learn the BN model, we apply techniques from statistical-relational learning, a recent field that combines AI and machine learning [21, 54, 16]. The *model log-likelihood ratio* (LR) is the log-ratio of the object model likelihood to the class model likelihood. This ratio quantifies how the probabilistic associations that hold in the general population deviate from the associations in the object data substructure. While the likelihood ratio discriminates relational outliers better than the class model likelihood alone, it can be improved further by applying two transformations: (1) a mutual information decomposition, and (2) replacing log-likelihood differences by log-likelihood distances. We refer to the resulting novel score as the *log-likelihood distance*.

**Evaluation** We analyze two real-world data sets, from the UK Premier League and the Internet Movie Database (IMDb). The empirical distributions of the ELD metric shows that the likelihood ratios highlight individual deviations more strongly. Success metrics, such as the Player's salary, provide an independent score for comparison with the ELD score. The data shows a surprisingly strong correlation between ELD and success metrics.

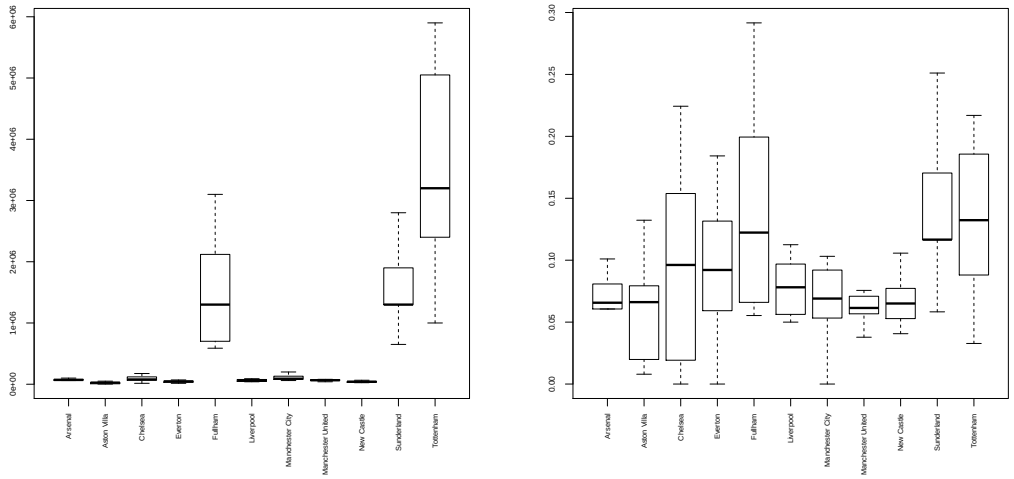
## 5.2 Preliminary Analysis

In this section we first study the relationship between several factors that we haven't included them in our modeling and think may have an effect on ranking individuals. We stored salary, nationality and age of 120 Players of Premier League in order to investigate effect of each of these factors on success of players.

**Fact #1: Some teams tend to pay more:** Figure 5.1(a) shows how different teams are in terms of their players salaries. Some teams have much larger budgets than others. Therefore, a player in "Tottenham" may not be necessarily performing substantially better than a player in the same position in "Manchester United", while there is substantial difference between the salaries of players in Tottenham and Manchester United. Fortunately, salaries of most of the teams in Premier League are in the same range and we can expect this fact to have little effect in the predicting of the success.

**Fact #2 Player's nationality has little effect on player's salary** Previous research on Soccer shows that nationality of the players sometimes affect the salary of the players regardless of their performance [60].
"The phrase 'Brazilian soccer player' is like the phrases 'French chef' or 'Tibetan monk.' The nationality expresses an authority, an innate vocation for the job-whatever the natural ability." [44]
We investigated effect of nationality on Players Salary in Premier League and showed that it has very little effect in our domain. Figure 5.3 shows the distribution of salaries across different nationalities.

**Fact #3 Older player from age 30-33 tend to have higher salaries** It is intuitive that Player's value increases as age increases (it has a peak for each player according to [57]). It takes time to accumulate fame and experience. Figure 3.6 shows the salary of the players




(a) Sum of Salary of the players of different team (b) Sum of Salary of the players of different team after normalization

Figure 5.1: Comparison of total salaries that different teams pay to their players

in different age. Please note this fact is not in favour of our method too. A famous player and a young player may play equally good but the famous player may have a higher salary due to his fame.

## 5.3 Related Work

Ranking individuals is a useful for many applications in Information Retrieval, Natural Language Processing and Data mining. In sports domain specifically ranking players is attractive and has many applications. Analyzing sports data can make significant difference in scoring players, signing contracts and preventing injuries. Pei *et al.* propose a reference based method that use relative degree of density with respect to a fixed set of reference points to calculate the neighbourhood density of a data point. They aim to find outstanding players based on two test settings: 1) games player, goals scored and shooting percentages. 2) points scored, plus minus statistics and penalty minus.

[57] *et al.* focused on the valuation of draft order in the SuperDraft. The valuation of draft order was first introduced in National Football League and proved to be very useful to the

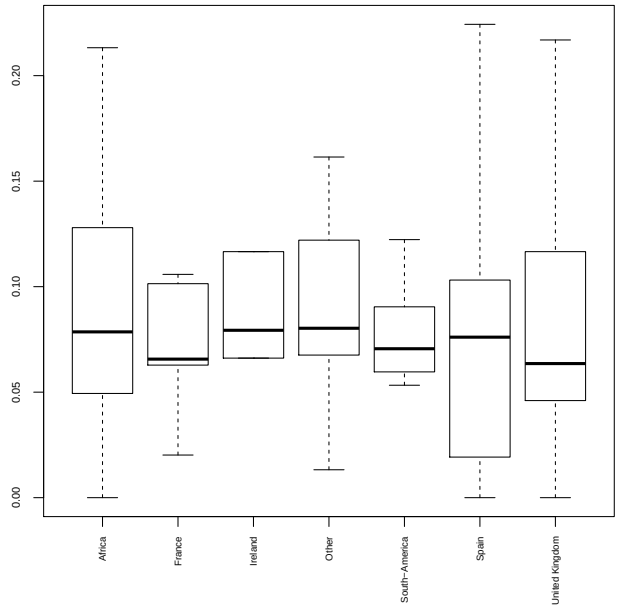



Figure 5.2: Salaries of players across different nationalities.

coach in trading players. They first estimate career trajectories of players and then assess the value of the draft position by introducing some performance measure. They used time played and salary of the player as ground truth in order to validate their method. [42] *et al.* developed an index to rate players regardless of their playing speciality based on player contribution to wining performances.

## 5.4 Result

The quality of generalization of population is the key to the performance of ELD metric in ranking individuals. For example, in premier league we expect most players to be in the range of good players. So the more different the player is from the population we can consider it as a sign of detecting exceptionally good players. Otherwise, it can be interpreted as the player can be both exceptionally good or exceptionally bad. Table 5.1,5.2 and 5.3 show the result of comparing the ELD of the individuals and their success metric.

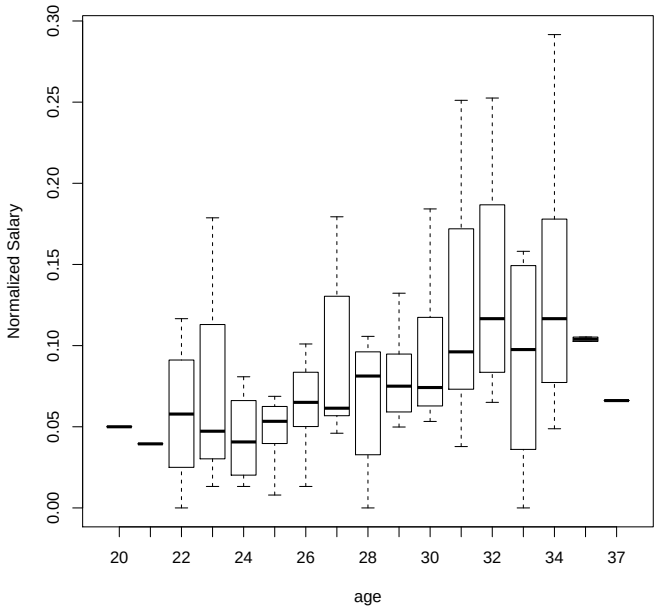



Figure 5.3: Salaries of players in different age.

| Metric | Sum of SavesMade | TimePlayed | Salary |
|---|---|---|---|
| ELD | 0.83 | 0.87 | 0.60 |

Table 5.1: Correlation between ELD metric and success metric of Goalies .

| Metric | Shots On Target | TimePlayed | Salary |
|---|---|---|---|
| ELD | 0.67 | 0.62 | 0.72 |

Table 5.2: Correlation between ELD metric and success metric of Strikers .

| Genre | Sum of Rating | Ranking |
|---|---|---|
| Action | 0.68 | 0.30 |
| Drama | 0.78 | 0.31 |
| Comedy | 0.85 | 0.42 |

Table 5.3: Correlation between ELD metric and success metric of Strikers .