Exploring the Kaggle European Soccer database with Bayesian Networks: the case of the Italian League Serie A

Maurizio Carpita and Silvia Golia

Abstract In the last decade, the application of statistical techniques to the sport field significantly increased. One of the most famous sports in the world is the soccer (football), and the present work deals with it, using data referred to the seasons from 2009/2010 to 2015/2016 of the Italian League Serie A extracted from the Kaggle European Soccer database. The players overall performance indicators, obtained on the basis of the players position or role (forward, midfielder, defender and goalkeeper), are used to predict the result of the matches by applying the Bayesian Networks as well as the Naive Bayes and the Binomial Logistic Regression models, considered as their competitors.
Abstract *Nell'ultima decade, l'applicazione di tecniche statistiche in ambito sportivo ha avuto un incremento significativo. In questo lavoro che riguarda il calcio, uno degli sport più famosi al mondo, si usano i dati relativi alle stagioni 2009/2010-2015/2016 del Campionato Italiano di Serie A estratti dal database Kaggle European Soccer. Gli indicatori di performance complessiva dei giocatori, ottenuti sulla base della loro posizione ovvero del loro ruolo (attaccante, centrocampista, difensore e portiere), sono utilizzati per prevedere il risultato delle partite utilizzando le Reti Bayesiane così come i modelli Naive Bayes e Logistico, considerati loro competitori.*

Key words: Kaggle European Soccer database, Bayesian Networks, Naive Bayes, Binomial Logistic Regression Model, Italian League Serie A

Maurizio Carpita
University of Brescia, Department of Economics and Management, C.da S.Chiara, 50 - 25122 Brescia, Italy, e-mail: maurizio.carpita@unibs.it

Silvia Golia
University of Brescia, Department of Economics and Management, C.da S.Chiara, 50 - 25122 Brescia, Italy, e-mail: silvia.golia@unibs.it

1 The Kaggle European Soccer database

In the last decade, the application of statistical techniques to the sport field and in particular to the European soccer (football) significantly increased, especially to predict matches' results [1, 15] using for example players performance statistics [10, 13] also for the Italian League [3, 4]. In the big data era, many databases are constructed and used to develop predictive models. On-line platforms for predictive modeling and analytics competitions, as Kaggle (www.kaggle.com), emerged and developed a meeting point for data scientists. This study uses data from the Kaggle European Soccer (KES) database [12], that contains data about 28,000 players and about 21,000 matches of the championship leagues of 10 countries and 7 seasons from 2009/2010 to 2015/2016, resulting in one of the biggest open database devoted to the soccer leagues of European countries, with data about players, teams and matches for several seasons [5].

The *Player Attributes table* of the KES database contains 33 variables, that represent player's performance indicators on the 0-100 scale with respect to overall and different abilities of the soccer play (power, mentality, skill, movement, attacking, goalkeeping), built with the experts classification of the EA Sports FIFA videogame. To develop the models of this study, only the players overall performance indicators and match result (in terms of goals scored by home and away teams) for the seasons from 2009/2010 to 2015/2016 of the Italian League Serie A are used. Given the presence of some missing values, the number of available matches is 2,587 instead of 2,660. The players overall performance indicator has been averaged based on the players' position or role, that is forward (FOR), midfielder (MID), defender (DEF) and goalkeeper (GOK), in each match according to the coach decisions before the match takes place. As explained in [5], the players' role is obtained from the *Match table* of the KES database, which contains X and Y coordinates representing the positions of the 22 players on the soccer pitch.

The aim of the study is twofold. First, one wants to explore these data using the Bayesian Networks (BN) as main model, to be compared to Naive Bayes (NB) and Binomial Logistic Regression (BLR) models, second, one wants to evaluate the power of overall performance indicators of the four roles in a soccer team (goalkeeper, defender, midfielder and forward role) in predicting the matches' results.

The paper is organized as follows. Sect. 2 contains a brief description of the theory underlying BN, NB and BLR models, whereas Sect. 3 reports the main results of the application of the three models to the data under study. Conclusions and ideas for future research follow in Sect. 4.

2 The three statistical models

Probabilistic networks are graphical models that explicit through a graph, the interactions among a set of variables represented as vertices or nodes of the graph. Any pair of unconnected nodes of the graph indicates (conditional) independence between the variables represented by these nodes under certain conditions that can be read from the graph itself. Hence, a probabilistic network captures a set of (conditional) dependence and independence properties associated with the variables represented in the network [8]. BN belong to the class of probabilistic networks. The underlined graph is called Directed Acyclic Graphs (DAG). A DAG $\mathscr{G}$ is a pair $\mathscr{G} = (\mathbf{V}, E)$, where $\mathbf{V}$ is a finite set of distinct vertices, $\mathbf{V} = \{V_i\}_{i=1}^{k}$, which correspond to a set of random variables $\mathscr{X} = \{X_{V_i}\}_{i=1}^{k}$ indexed by $\mathbf{V}$, $E \subseteq \mathbf{V} \times \mathbf{V}$ is the set of directed links (or edges) between pairs of nodes in $\mathbf{V}$. An ordered pair $(V_i, V_j) \in E$ denotes a directed edge ($\rightarrow$) from node V_i to node V_j; V_i is said to be a parent of V_j whereas V_j a child of V_i. The set of parents of a node V shall be denoted by $pa(V)$. A Bayesian Network (BN) over the variables $\mathscr{X}$ is defined as the triplet $\mathscr{N} = (\mathscr{X}, \mathscr{G}, \mathscr{P})$, where $\mathscr{G}$ is a DAG and $\mathscr{P}$ is a set of conditional probability distributions containing one distribution $P(X_v|X_{pa(v)})$ for each random variable $X_v \in \mathscr{X}$, where $X_{pa(v)}$ denotes the set of parent variables of variable X_v. The joint probability distribution $P(\mathscr{X})$ over the set of variables $\mathscr{X}$ is factorized as

$$P(\mathscr{X}) = \prod_{v \in \mathbf{V}} P(X_v|X_{pa(v)}). \tag{1}$$

So, a BN can be described in terms of a qualitative component, that is the DAG, and a quantitative component, consisting of the joint probability distribution (1).

The construction of a BN runs in two steps. First, one identifies the interactions among the variables generating a DAG, then the joint probability distribution has to be specified in terms of the set of conditional probability distributions $P(X_v|X_{pa(v)})$.

The DAG can be derived either manually or automatically from data. In order to automatically find the BN structure, several algorithms have been proposed in the literature, falling under three broad categories: constraint-based, score-based, and hybrid algorithms. Constraint-based algorithms make use of conditional independence tests focusing on the presence of individual arcs, score-based algorithms assign scores to evaluate DAGs as a whole, whereas hybrid algorithms combine constraint-based and score-based algorithms. The method used in this paper to derive the DAG is the Hill Climbing (HC), implemented in the R package `bnlearn` provided by Scutari [16]. It belongs to the class of the score-based algorithms and it consists in exploring the search space starting from an empty DAG and adding, deleting or reversing one arc at a time until the score considered can no longer be improved [17].

Given a BN, it can be used to answer questions (queries) related to the domain of the data that goes beyond the description of its behavior; for BN the process of answering questions is also known as probabilistic reasoning or belief updating. Basically, it focuses on the calculus of the posterior probabilities given a new piece of information called evidence. In fact, suppose to have learned a BN, it is used to compute the effect of new evidence Ev on one or more target variables X' using the knowledge encoded in the BN, that is to compute posterior distribution $P(X'|Ev)$. The procedure used in the paper makes use of the junction tree representation of a BN which is composed by cliques (a clique is a maximal complete subgraph of a moralization of the DAG $\mathscr{G}$). Belief updates can be performed efficiently using Kim and Pearls Message Passing algorithm [9].

Even if all the variables play the same role in the construction and usage of the BN, in the present paper one of the variables will be considered as target variable, similar to the variable Y used in the definition of the two competing models that follows.

In order to compare the performance of BN with the performances of competing models, the NB and BLR models are taken into account.

The Naive Bayes (NB) is one of the simplest restricted probabilistic graphical models [6]. It is characterized by a structure where one single class variable is parent of the remaining attribute variables, which are conditionally independent given the class variable. So, let Y be the class variable and let $\{X_i\}_{i=1}^{p}$ be the attribute variables, their joint probability is factorized as $P(Y, X_1, \cdots, X_p) = P(Y) \cdot \prod_{i=1}^{p} P(X_i|Y)$. From the definition of conditional probability, $P(Y|X_1, \cdots, X_p)$ is given by the following expression:

$$P(Y|X_1, \cdots, X_p) = \alpha \cdot P(Y) \cdot \prod_{i=1}^{p} P(X_i|Y), \tag{2}$$

where α is a normalization constant. Equation (2) is the common definition of a NB.

The Binomial Logistic Regression (BLR) model belongs to the family of the generalized linear models and it is used to estimate the probabilities of the two categories of a binary dependent variable Y using a set of covariates or predictors $\mathbf{X} = \{X_i\}_{i=1}^{p}$ [7]. The model assumes that the binary response variable is distributed as a Bernoulli with probability π and a logit link function. It can be expressed through the *logit transformation* as:

$$logit[\pi(\mathbf{X})] = log\left[\frac{\pi(\mathbf{X})}{1-\pi(\mathbf{X})}\right] = \mathbf{X}\boldsymbol{\beta} \tag{3}$$

where $\pi(\mathbf{X}) = P(Y = 1|\mathbf{X})$ and $\boldsymbol{\beta}$ is the vector of coefficients involved in the linear predictor $\mathbf{X}\boldsymbol{\beta}$. The BLR model was already applied to predict the result of soccer matches in other studies [2, 5, 11].

3 Statistical evidences for the Italian League Serie A

As explained in Sect. 1, the dataset used in this paper was obtained from the KES database and contains the overall performance indicators for the four roles in a soccer team (goalkeeper, defender, midfielder and forward role), used as predictors, and the matches' results reported in terms of goals scored by the home and away teams. The number of overall performance indicators is eight and corresponds to p for the NB and BLR models. From the goals scored by the two teams of the match, it is possible to determine the outcome of the match from the home team point of view, *result*, classified as win, draw and loss. Preliminary analysis, using this classification for the *result* variable, has shown a high difficulty of the BN in predicting the draw with respect to win and loss; similar performances were observed in previous studies [3, 4]. This finding could be due to the fact that a draw is an outcome characterized by a higher degree of uncertainty, as well as to class imbalance [14] as in this case, where percentage of wins is 46.1%, whereas percentages of loss and draw are 26.4% and 27.5% respectively. In order to partially overcome this problem, the *result* variable was dichotomized into two categories: WIN and NOWIN (that is loss or draw) of the home team. Clearly, a model that takes into account three categories for the outcome with satisfactory prediction accuracy should be preferred to a model that involves less categories. Nevertheless, a model based on a binary variable, as the dichotomized *result*, can be of interest for the home team (obviously interested to the WIN probability) as well as the away team (interested to the NOWIN probability of the home team, that is the NOLOSS probability of the away team). The *result* variable represent the target variable denoted by Y in the NB and BLR models. So, the variables used in all the analyses are nine and this number corresponds to k for BN; the link between p and k is the following: $k = p+1$.

Given that the dataset under study contains continuous variables (averages by players role of the overall performance indicator on 0-100 scale) and a discrete variable (outcome of the match), it is necessary to discretize the continuous variables. In applying BN it is possible to manage hybrid database like the one under study, but it is necessary to constraint the relation parent-child, imposing that a discrete variable may only have discrete parents [8]. In the context of this paper, this kind of constraint implies that the overall performance indicators can not be parents of the outcome of the match and this constraint appears unrealistic.

The method used to discretize the players' performance indicators, is the Equal Frequency Discretizer that involves the quantiles of the variable's distribution. The number of categories was fixed to four.

In order to identify the best combination of score-based algorithm and score that gives a DAG with highest score, a cross-validation has been performed. The scores under evaluation were the Bayesian Information criterion (BIC) and the Bayesian Dirichlet Equivalent (BDE) uniform posterior probability of the DAG associated with a uniform prior over both the space of the DAGs and of the parameters [17]. The selected score was the BDE with imaginary sample size (iss) equal to 100; iss determines how much weight is assigned to the prior distribution compared to the data when computing the posterior. Fig. 1 shows the obtained DAG. It can be seen that the home and away teams are well separated and with a coherent structure of links. In fact, the goalkeeper role is connected with the defender and midfielder roles, the defender role is linked to the midfielder role and the forward role is related to the midfielder and defender roles. Moreover, the structure of relations between the performance indicators of both the home and away teams is the same, whereas the variables with a direct link with the variable *result* are different. For the home team the role directed linked to the variable *result* is the midfielder role whereas the one of the away team is the defender role. This finding is coherent with the match strategies chosen by the two coaches: generally, the home team goal is to win the match, so the midfielder role is important, whereas the away team goal is not to lose the match and in this case, the defender role is the strategic one.

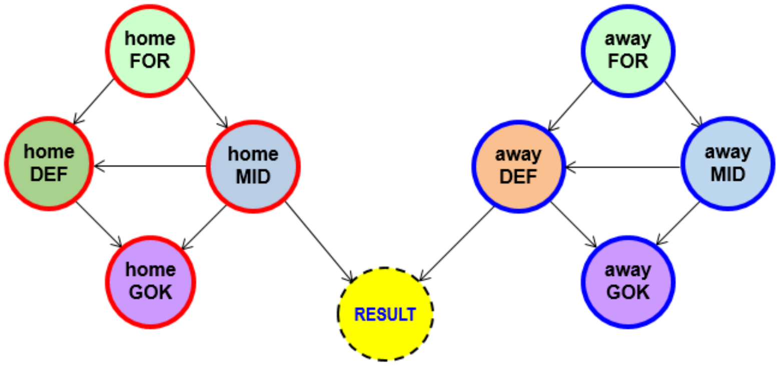

Fig. 1 DAG for the Italian League Serie A, Seasons 2009/2010-2015/2016

In order to compare the performance of the BN applied to the data, NB and BLR models were estimated. Table 1 reports the estimates of the parameters involved in the BLR model with the corresponding z-statistic. The performance indicators are maintained categorical in order to use the same data as the BN; same comments arise from the analysis of the results of the BLR model with continuous performance indicators. The most significant variables for the BLR model are *home_MID* and *away_DEF*, which are the ones with directed links to *result* in BN.

Table 1 Parameter estimates and z-statistics for the BLR Model

Perf. Indicator	$\hat{\beta}$	z-stat	Perf. Indicator	$\hat{\beta}$	z-stat
home_FOR2	-0.023	-0.181	away_FOR2	-0.094	-0.793
home_FOR3	0.029	0.206	away_FOR3	-0.021	-1.157
home_FOR4	0.319	1.959	away_FOR4	-0.289	-1.758
home_MID2	0.155	1.231	away_MID2	0.057	0.463
home_MID3	0.555	3.817	away_MID3	0.077	0.527
home_MID4	0.776	4.173	away_MID4	-0.466	-2.430
home_DEF2	0.086	0.693	away_DEF2	-0.260	-2.165
home_DEF3	0.336	2.286	away_DEF3	-0.420	-2.788
home_DEF4	0.478	2.689	away_DEF4	-0.672	-3.692
home_GOK2	-0.035	-0.288	away_GOK2	-0.078	-0.649
home_GOK3	-0.007	-0.060	away_GOK3	-0.153	-1.309
home_GOK4	-0.013	-0.087	away_GOK4	0.010	0.072

All the methods considered calculate the probability of win given new information. In order to predict the result of the match in terms of WIN-NOWIN of the home team, the simple majority rule is the most popular way to convert a probability into a predicted result; this is the rule used in the paper. Precision of model predictions can be evaluated by computing some indexes such as the *Accuracy*, which expresses how effectively the model predicts matches' results, the *Sensitivity*, which expresses how effectively the model predicts WIN matches and the *Specificity*, which expresses how effectively the model predicts NOWIN matches. Moreover, the accuracy of a model can be compared with the so called *null accuracy*, which measures the accuracy obtained without models and corresponds to the highest observed frequency of the two possible results of the match. Table 2 reports the prediction performances of BN, NB and BLR considering 500 random samples of 500 matches form the 2,587 available. For each sample of 500 matches, the DAG is the one reported in Fig. 1, whereas in applying NB and BLR, all the overall performance indicators are taken into account as predicting variables of *result*. All the conditional probabilities involved in BN and NB and the coefficients of BLR are computed making use of the remaining 2,087 matches.

Table 2 Mean predictive capability of the models based on the prediction of 500 matches' results randomly sampled 500 times (standard errors are in parenthesis)

Model	Accuracy	Sensitivity	Specificity
BN	0.623 (0.020)	0.550 (0.055)	0.688 (0.052)
NB	0.632 (0.021)	0.577 (0.032)	0.679 (0.030)
BLR	0.630 (0.021)	0.525 (0.031)	0.721 (0.031)

All methods have similar accuracy (about 63%), which is about ten percentage points greater than the null accuracy (53.9%). Moreover, BN and NB are more accurate in the WIN prediction (sensitivity) than BLR, which is more accurate in predicting NOWIN (specificity). Finally, standard errors for sensitivity and specificity of BN are greater than those of NB and BLR of about two percentage points.

4 Conclusions and future research

The paper shows some results obtained analyzing data regarding the Italian League Serie A extracted from the KES database; the aim was to explore these data using the BN and evaluate the power of overall performance indicators of the four roles in a soccer team (goalkeeper, defender, midfielder and forward role) in predicting the matches' results. A preliminary analysis suggested the necessity to collapse the match's results loss and draw in a unique category. Moreover, due to the fact that BN works with discrete variables, the overall performance indicators were discretized. In addition to BN, NB and BLR models were considered as competitors. All the three models have similar accuracy (around 0.63) significantly greater than null accuracy (0.54), so it is possible to conclude that the variables considered have some predictive power. Regarding the application of BN, the DAG expressed by the data shows coherent structure of links between the roles and maintains the home and away teams indicators well separated. Moreover the variables with a direct link with the variable *result* were the midfielder role for the home team and the defense role for the away team. This finding is coherent with the match strategies chosen by the two coaches: generally, the home team goal is to win the match, so the midfielder role is important, whereas the away team goal is not to lose the match and in this case, the defender role is the strategic one.

This study can be extended in various directions. For example, an advantage in using BN with respect of other models is the possibility to use partial information to predict a match's result, and this will be argument of future research. Moreover, another interesting development of this paper will be to extend the analysis to leagues of other countries in Europe in order to verify similarities in the results.

References

1. Albert, J., Glickman, M.E., Swartz, T.B., and Koning, R.H.: Handbook of Statistical Methods and Analyses in Sports. CRC Press (2017)
2. Alves, A.M., Mello, J.C.C.B.S., Ramos, T.G., Sant'Anna, A.P., et al.: Logit models for the probability of winning football games. Pesquisa Operacional **31(3)**, 459–465 (2011) doi: 10.1590/S0101-74382011000300003
3. Carpita, M., Sandri, M., Simonetto, A., Zuccolotto, P.: Football mining with R. In: Yanchang, Z., Yonghua, C. (eds.) Data Mining Applications with R, pp. 397-433. Springer (2014)
4. Carpita, M., Sandri, M., Simonetto, A., Zuccolotto, P.: Discovering the drivers of football match outcomes with data mining. Quality Technology & Quantitative Management **12(4)**, 561–577 (2015) doi: 10.1080/16843703.2015.11673436
5. Carpita M., Ciavolino E., Pasca P.: Exploring and modelling team performances of the Kaggle European Soccer Database. Submitted (2018)
6. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine Learning **29**, 131–163 (1997) doi: 10.1023/A:1007465528199
7. Hosmer, D.W.Jr, Lemeshow, S., Sturdivant, R.X.: Applied logistic regression, 3rd ed. John Wiley & Sons (2013)
8. Kjærulff, U.B., Madsen, A.L.: Bayesian networks and influence diagrams: a guide to construction and analysis. Springer, New York (2013)
9. Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. MIT Press (2009)
10. Leung, C.K., Joseph, K.W.: Sports data mining: predicting results for the college football games. Procedia Computer Science **35**, 710–719 (2014) doi: 10.1016/j.procs.2014.08.153
11. Magel, R., Melnykov, Y.: Examining influential factors and predicting outcomes in European soccer games. International Journal of Sports Science **4(3)**, 91–96 (2014) doi: 10.5923/j.sports.20140403.03
12. Mathien, H.: European Soccer Database. (2016) Data retrieved from
http://www.kaggle.com/hugomathien/soccer
13. McHale, I.G., Szczepański, L.: A mixed effects model for identifying goal scoring ability of footballers. Journal of the Royal Statistical Society: Series A **177(2)**, 397–417 (2014) doi: 10.1111/rssa.12015
14. Menardi, G., Torelli, N.: Training and assessing classification rules with imbalanced data. Data Mining and Knowledge Discovery **28(1)**, 92–122 (2014) doi: 10.1007/s10618-012-0295-5
15. Odachowski, K., Grekow, J.: Using bookmaker odds to predict the final result of football matches. In Graña, M., Toro, C., Howlett, R.J., Jain, L.C., (eds.) Knowledge Engineering, Machine Learning and Lattice Computing with Applications, pp. 196-205. Springer, Berlin Heidelberg (2013)
16. Scutari, M.: Learning bayesian networks with the bnlearn R package. Journal of Statistical Software **35(3)**, 1–22 (2010) doi: 10.18637/jss.v035.i03
17. Scutari, M., Denis J-B.: Bayesian Networks With Examples in R. CRC Press, Taylor & Francis Group (2015)