The European Conference
on Machine Learning and Principles
and Practice of Knowledge Discovery in Databases

ECML/PKDD 2014
PhD Session Proceedings

September 15-19, 2014
Nancy, France

Edited by

Radim Belohlavek
Bruno Crémilleux

Program Committee

Program Committee Co-Chairs

Radim Belohlavek	(Palacky University, CZ)
Bruno Crémilleux	(University of Caen, FR)

Program Committee Members

Stephane Canu	(INSA Rouen, FR)
Peggy Cellier	(INSA Rennes, FR)
Sanjay Chawla	(University of Sydney, AU)
Carlotta Domeniconi	(George Mason University, VA, US)
Antoine Doucet	(University of Caen, FR)
Richard Emilion	(University of Orleans, FR)
Georgiana Ifrim	(University College Dublin, IE)
Dino Ienco	(IRSTEA, FR)
François Jacquenet	(University Jean Monnet, Saint-Etienne, FR)
Jiri Klema	(Czech Technical University, Prague, CZ)
Rudolf Kruse	(Otto von Guericke University Magdeburg, DE)
Marzena Kryszkiewicz	(Warsaw University of Technology, PL)
Sergei Kuznetsov	(National Research University HSE, Moscow, RU)
Donato Malerba	(University of Bari, IT)
Pauli Miettinen	(Max-Planck-Institut fr Informatik, DE)
Amedeo Napoli	(CNRS & INRIA Nancy Grand Est)
Sergei Obiedkov	(National Research University HSE, Moscow, RU)
Panagiotis Papapetrou	(Stockholm University, SE)
Chedy Rassi	(INRIA Nancy Grand Est)
Sebastian Rudolph	(University of Karlsruhe, DE)
Dan Simovici	(University of Massachusetts, Boston, MA, US)
Arnaud Soulet	(University Francois Rabelais of Tours, FR)
Gerd Stumme	(University of Kassel, DE)

Table of Contents

Papers with Oral Presentations

Search for User-related Features in Matrix Factorization-based Recommender Systems 1
Marharyta Aleksandrova, Armelle Brun, Anne Boyer and Oleg Chertov

Optimistic Active Learning for Classification 11
Timothé Collet and Olivier Pietquin

Recommender-based Multiple Classifier System 21
Yury Kashnitsky

Clustering Boolean Tensors 31
Saskia Metzler and Pauli Miettinen

On Improving Operational Planning and Control in Public Transportation Networks using Streaming Data: A Machine Learning Approach 41
Luis Moreira-Matias, João Mendes-Moreira, João Gama and Michel Ferreira

Inference of Switched Biochemical Reaction Networks Using Sparse Bayesian Learning 51
Wei Pan, Ye Yuan, Aivar Sootla and Guy-Bart Stan

Heterogeneous Bayes Filters with Sparse Bayesian Models: Application to state estimation in robotics 61
Alexandre Ravet and Simon Lacroix

Be In The Know: Connecting News Articles to Relevant Twitter Conversations 71
Bichen Shi, Georgiana Ifrim and Neil Hurley

Papers with Poster Presentations

Evaluating Collaborative Filtering: Methods within a Binary Purchase Setting 81
Stijn Geuens, Kristof Coussement and Koen W. De Bock

An opinion mining Partial Least Square Path Modeling for football betting . . 91
Mohamed El Hamdaoui and Jean-Valère Cossu

Parallel Learning Algorithm for Large-Scale Regression with Additive Models . 101
Valeriy Khakhutskyy and Markus Hegland

Generalizing, Optimizing, and Decoding Support Vector Machine Classification 111
Mario Michael Krell, Sirko Straube, Hendrik Wöhrle and Frank Kirchner

Robust Optimization using Machine Learning for Uncertainty Sets 121
Theja Tulabandhula and Cynthia Rudin

Heterogeneous Dataflow Hardware Accelerators for Machine Learning on Reconfigurable Hardware 129
Hendrik Woehrle, Johannes Teiwes, Mario Michael Krell, Anett Seeland, Elsa Andrea Kirchner and Frank Kirchner

Multivariate Normal Distribution Based Multi-Armed Bandits Pareto Algorithm 139
Saba Q. Yahyaa, Madalina M. Drugan and Bernard Manderick

Managing Ventilation Systems for Improving User Comfort in Smart Buildings using Reinforcement Learning Agents 149
Jiawei Zhu, Fabrice Lauri, Abderrafiaa Koukam and Vincent Hilaire

A Framework for Pattern Classifier Selection and Fusion 159
Fabio A. Faria, Anderson Rocha and Ricardo da S. Torres

Affinity Analysis between Researchers using Text Mining and Differential Analysis of Graphs 169
Luís Trigo and Pavel Brazdil

NASSAU: Description Length Minimization for Boolean Matrix Factorization . 177
Sanjar Karaev

Preface

The present volume contains papers accepted for oral and poster presentation in the PhD Session of the 2014 EMCL/PKDD conference held September 15-19, 2014 in Nancy, France. The objective of the session was to provide an environment for students to exchange their ideas and experiences with peers and to get constructive feedback from senior researchers in machine learning, data mining and related areas. The topics for discussion were mainly ideas of students and their ongoing work in preparation of their PhD dissertations in machine learning and data mining.

The PhD session received 23 initial submissions from which 8 were finally accepted for oral presentation and 11for poster presentation. All submissions were reviewed typically by three reviewers on the basis of their originality, quality, significance, and presentation. The program of the PhD session consisted of oral and poster presentations of accepted papers and an invited talk by Bart Goethals (University of Antwerp, Belgium).

We would like to express our thanks to the authors who submitted their papers, to the invited speaker, and to the members of the Program Committee, who all helped make the PhD Session a successful event. We also thank Martin Trnecka (Palacky University, Olomouc) for his help in preparing this proceedings. Finally, we wish to thank the local organization team of ECML/PKDD 2014, and in particular Amedeo Napoli and Chedy Raïssi.

September 2014

Radim Belohlavek and Bruno Crémilleux
ECML/PKDD 2014 PhD Session Co-Chairs

An opinion mining Partial Least Square Path Modeling for football betting

Mohamed El Hamdaoui, Jean-Valère Cossu

LIA/Université d'Avignon et des Pays de Vaucluse
** 39 chemin des Meinajaries, Agroparc BP 91228, 84911 Avignon cedex 9, France
firstname.name@alumni.univ-avignon.fr

Abstract. In the last few years, football betting had known a large expansion in the world, using different ways to try to guess and predict the unknown in the sport. Every time, people try to prognosticate the results of matches using probabilistic, statistic and other methods to get the maximum benefits, especially with the emerging of betting websites. In this paper, we present an alternative approach, to state of the art probabilistic models, based on Partial Least Square Path Modeling (PLS-PM). We first show that the simple PLS model containing only statistical resources about each team are efficient to predict the team ranking at d+1 and this gives a state of the art prediction of match outcomes. We then take advantage of PLS ability of integrating complex and heterogeneous data to reach a practical model by including textual data, taken from tweets related to teams, that we previously classify by polarity using robust sentiment analysis in multiple languages. Another learning of our experiment is the role of the inner model in PLS when used for prediction purpose. Unlikely Bayesian networks, the latent variable used in the prediction need to be deeply inside the inner model and not considered as marginal outcomes, this to allow back and forth retro-propagation from multiple types of data. The main purpose of our work is to show that PLS-PM can be surprisingly efficient in predicting tournament outcomes for which temporal statistics and social network data are available if inference is based on central inner latent variables.

1 Introduction

Football is one of the most famous sport in the world, where betting on results is very popular. But it is not as easy as it looks, because even the football experts are not expert in prognostication as shown by [3]. Almost all systems and offices of betting use the probabilistic model to calculate odds linked to each part of bet. [5] gives in "Statistica Neerlandica", an example how scores are obtained using Poisson goals distribution. We want to prove, by these experiments, that probabilistic model is not the only way that allows to get efficient prognostications and intend to explore if a betting system based on correlation analysis of multiple and sparse data can be improved using different Partial Least Squares

** http://lia.univ-avignon.fr/

Path Modeling (PLS-PM). The remainder of the paper is structured as follows. First we present a simple model, based only on the ranking of teams. Then, we will use a model based on a few variables before combining these two solutions, and we will compare it with a probabilistic model, which is the rating used in different sport betting game. Finally, we will try to improve our PLS-PM model using text mining over twitter. Our work focus on on the leagues of France "League 1", England "Premier League", Spain "La Liga", and Italy "Serie A", each league is composed by 20 teams, and data for the first experimentation were obtained from the 14th, until the 23rd day of the league, and from the last 6 days for the model that contains textual data.

2 PLS-PM models

PLS-PM is a statistical method that allows studying and modeling complex relationships between observed (manifest) and latent variables. Data is analyzed like a structure made of blocks of manifest variables, and each block is summarized by a latent variable. This approach was developed by Herman WOLD during the 70s of the last century, when he presented PLS for the first time in 1979. But its popularity just started recently to increase in different domains. PLS-PM is formally represented by two sets of linear equation, the inner model and the outer one. The first model represents the relationships between latent variables, while the other model represents the relationships between a latent variable and its manifest variables. PLS-PM is a way to estimate parameters, it is used to find complex linear regressions, based on the latent and manifest variables, by calculating the solution of the general underlying model of multivariate PLS. For more details on how to manipulate manifest variables and these relations we refer the reader to [1] and [7]. We try to use different models to demonstrate the ability of PLS-PM concerning prognostication on football games. By these experiments we want to find the success indicator (SI) of each team. This will allow us to compare two teams and then, prognosticate which one will win. We tried to combine in one hand the method used in [6] based on simple statistics concerning number of goals scored and conceded, and on the other hand, the method used in [5], where he makes the difference between matches played at home and away. This way, we had to duplicate our PLS-PM model, in order to calculate both at home and away SI. Before starting, let us show you the type of data that we use, it is a table containing information for each team (in table 1).

Table 1. Resume of the 23rd day of league 1

Team	GSH	GSA	PSH	PSA	GCH	GCA	PCH	PCA	WH	WA	Rank
PSG	35	19	1.00	0.91	-5	-10	0.67	0.27	9	7	20
ASMonaco	19	19	0.91	0.92	-6	-10	0.54	0.42	8	6	19
Lille	16	8	0.91	0.50	-6	-8	0.64	0.58	8	4	18

This table shows some statistics on the top 3 ranked teams, collected after crawling and scraping some football websites.
GS (H/A): Number of goals scored at home (H) or away (A).
PS (H/A): Percentage of game where the team scored goal(s).
GC (H/A): Number of goals conceded.
PC (H/A): Percentage of game where the team conceded goal(s).
WM (H/A): Number of won games.

2.1 Model based on Ranking

We started by an inner model that contains three latent variables: Attack, Defense, and Success. This first experiment consists in realizing a baseline model, including only the ranking of the league as a manifest variable, which reflects the latent variable Success (as shown in figure 1). Keep in mind that we used that model twice respectively for away and at home Success Index (SI).

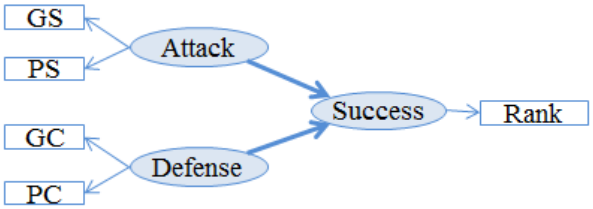

Fig. 1. Simple model based on Rank

Figure 1 represents the relation that exists between Attack, Defense, and Success, and our model is based on how each variable impacts other variables, so we can express our model in the following equation.

$$\text{Success}_{rank} = f(\text{Attack}, \text{Defense})$$

2.2 Model based on won matches

In this second experiment, we replaced the variable Ranking, by other latent variables like number of wins (as shown in figure 2).

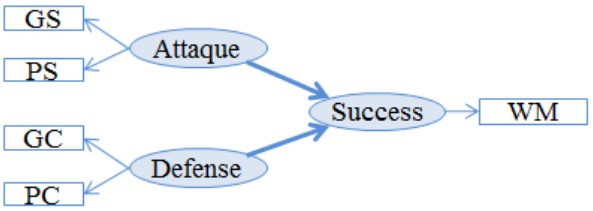

Fig. 2. Simple model based on number of won matches

In figure 2, Success is based this time on number of won games, so we can express it like:

$$\text{Success}_{won} = f(\text{Attack}, \text{Defense})$$

2.3 Model based on Ranking and won matches

The next experiment consists in mixing the previous models, to get this PLS-PM model (as described in figure 3):

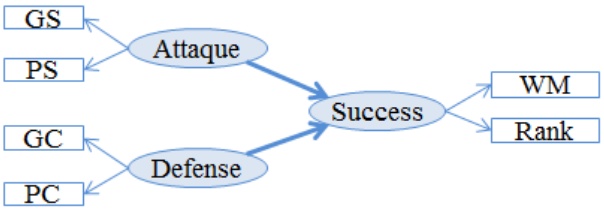

Fig. 3. PLS-PM model based on number of won matches and Rank

$$\text{Success}_{(rank,won)} = f(\text{Attack}, \text{Defense})$$

In each experiment, we used our results to verify the efficiency of the model, by prognosticating the results of matches, and we got results that gives the number of the right prognostication in ten games by day (figure 4).

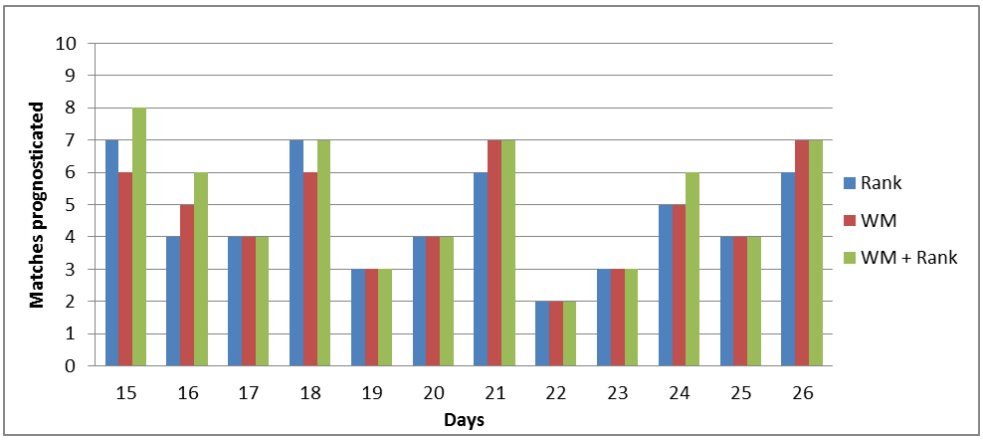

Fig. 4. Comparison between the previous models

Figure 4 compares the number of matches prognosticated by each model, from the 15^{th} week, until the 26^{th} one, it shows how, by using the model that combines ranking and won games, the number of correct matches prognosticated is greater than using models separately.

$$\text{Success day}_i(\text{WM} + \text{Rank}) \geq Max(\text{Success day}_i(\text{WM}), \text{Success day}_i(\text{Rank}))$$

As an interesting result it should be important to notice that the lowest values of correct matches prognosticated in one day is mostly due to the abundance of draws, which are difficult to predict (e.g. 22^{nd} week, there were 5 draws, which explain that our model predicted only 2 matches). So, we see that we can improve our prognostications by adding more manifest variables, such as number of points obtained in the last matches, as well as systems and websites of betting do based on previous matches.

2.4 Probabilistic model

Betting systems consider three values for each bet as it is represented in the following table (in table 2). Here, we consider the prognostication true if the result was a victory for Monaco, (1) means victory of the host team, because it had the highest probability of winning, but unluckily, this prognostication was false because the game finished by a draw which its odd was 3.34. Conversely, the prognostication in the second example was true, because Lyon, which has the lowest odd, won that game. So we had follow this way to calculate the results of other matches.

Table 2. Example for probabilistic model based on Odds

Host	Result	Visitor	(1)	(x)	(2)
Monaco	1 - 1	Lille	1.88	3.34	4.35
Nice	0 - 1	Lyon	4.48	3.54	1.82

(1): Host wins, (x): Draw, (2): visitor wins

Notice that odds are the inverse of probabilities values.

$$\text{Odds} = \frac{1}{p(\omega)}, \omega \in \Omega\{1, x, 2\}$$

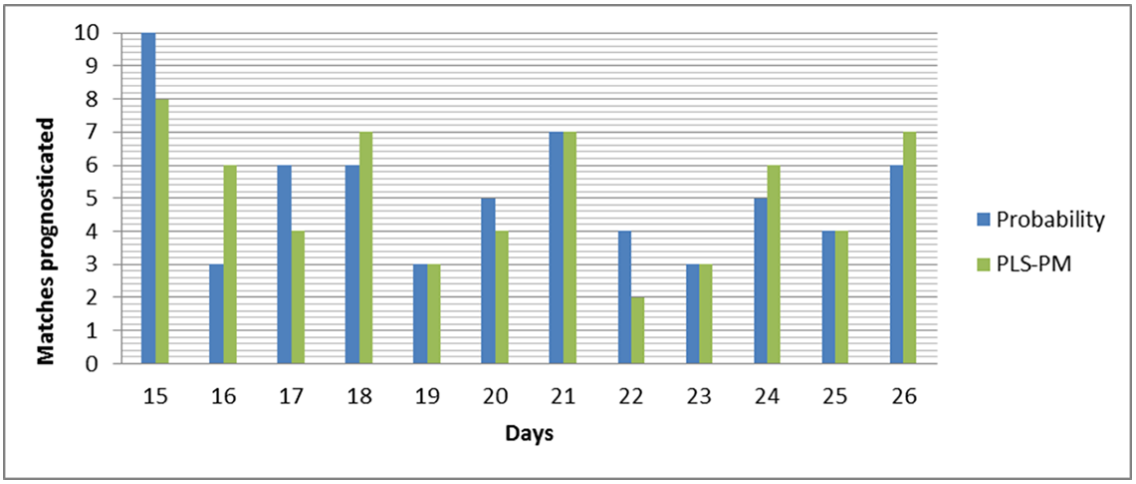

Fig. 5. Comparison between PLS-PM and probabilistic model

Figure 5 compares the probabilistic model, which is the state of the art, with our last PLS-PM model which assembles ranking and number of games won. These statistics between each model, prove how our PLS-PM model as efficient as the probabilistic model.

2.5 PLSPM with Twitter

The most interesting feature in PLS-PM, is that it can deal we a large variety of heterogeneous variables, provided that the correct model is set. Our next experiment consists in adding textual data in the model. As example we took the

reputation of each team on twitter, using twitteR package. It allows to collect tweets concerning each team. Then we used sentiment package in order to perform a 3-valued (positive, negative and neutral classes) polarity classification of these tweets using multiple languages opinion lexicon [2] [4]. As we mentioned, the difficulty relies in setting the correct model. For example, in a first trial we used a model which considers the reputation of a team as an element of its success (as described in figure 6). This first model downgraded all results as shown in (table 3).

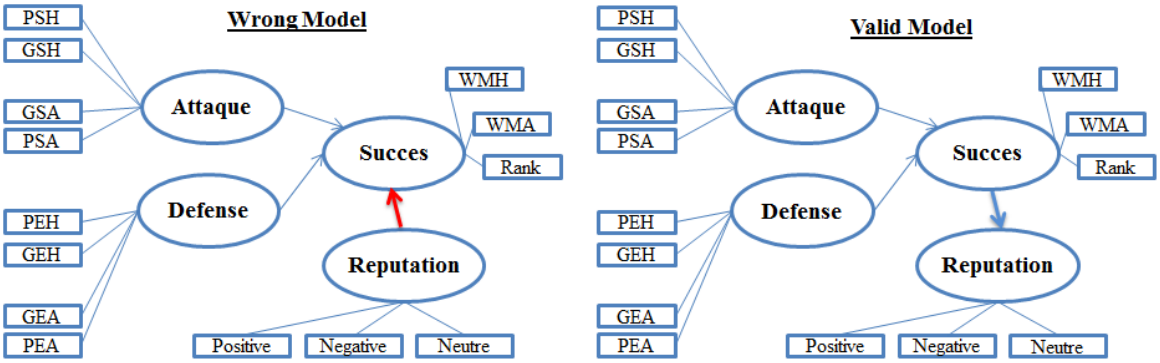

Fig. 6. The importance of choosing the right sense of relation between variables

Table 3. Result of wrong Model

Team	Success H	Success A
PSG	-1.7766	-1.7766
ASMonaco	0.8521	0.8521
Lille	-0.7359	-0.7359
OM	-0.3059	-0.3059
StdReims	2.2199	2.2199

In fact, it is not the reputation which affects success, but the opposite, so we changed the path matrix of the last model as described in figure 6.

$$\begin{cases} \text{Success} = f(\text{Attack}, \text{Defense}) \\ \text{Success} = f^{-1}(\text{Reputation}) \end{cases}$$

By next, we compared the result obtained by the probabilistic model with the improved PLS-PM model, and both are closer (table 4).

Figure 7 summarizes the last table, where we see that the difference between the number of matches predicted by probabilistic model, and PLS-PM model including text data (Twitter) is only 1 match in a total of 240, and 9 matches more than our basic PLS-PM model. This is an encouraging result, knowing that Probabilistic model, based on Poisson distribution, has a high performance of prediction because it integrates time series.

Table 4. Number of match predicted by each model

	Games	Results of matches	Model	Predictions	%
FR	70	1 - 30	Probability	38	0,54
	70	N - 19	PLSPM	33	0,47
	70	2 - 21	Twitter	37	0,53
ES	60	1 - 29	Probability	30	0,50
	60	N - 16	PLSPM	28	0,47
	60	2 - 15	Twitter	28	0,47
EN	50	1 - 23	Probability	30	0,60
	50	N - 8	PLSPM	27	0,54
	50	2 - 19	Twitter	29	0,58
IT	60	1 - 29	Probability	35	0,58
	60	N - 12	PLSPM	36	0,60
	60	2 - 19	Twitter	38	0,63

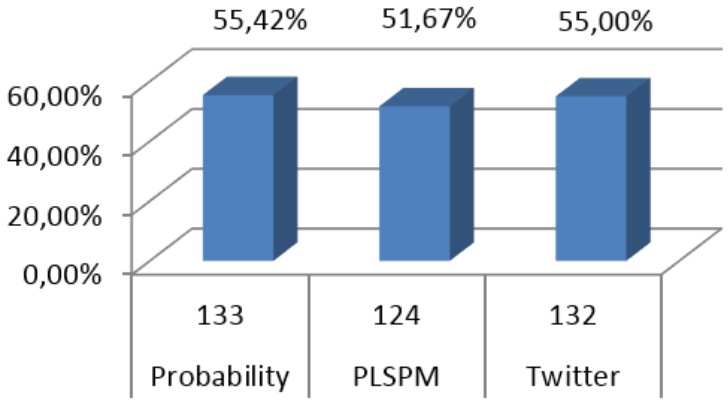

Fig. 7. Percentage of match predicted by each model

The most important thing to notice, concerning PLS-PM model including text data, is that it improves the number of predicted drawn matches, and this is what explains how it outperforms PSL-PM basic model, especially in the case of teams with similar rankings. The method considering that a game will finish by a draw is resumed in the next formula :

$$\text{Draw} \Leftrightarrow \|\text{Success at Home}–\text{Success Away}\| \leq 0.02$$

Tweets give in fact the latest information concerning one team like when a player has been being injured, or excluded and would not play next game. In addition, it is significant to know if a team is well supported or not, or it is in a good financial situation.

3 PLS-PM and Bootstrapping

Our last experiment consisted in comparing 3 PLS-PM models (as described in figure 8), where we applied the bootstrapping method to obtain information about the variability of our different estimated variables.

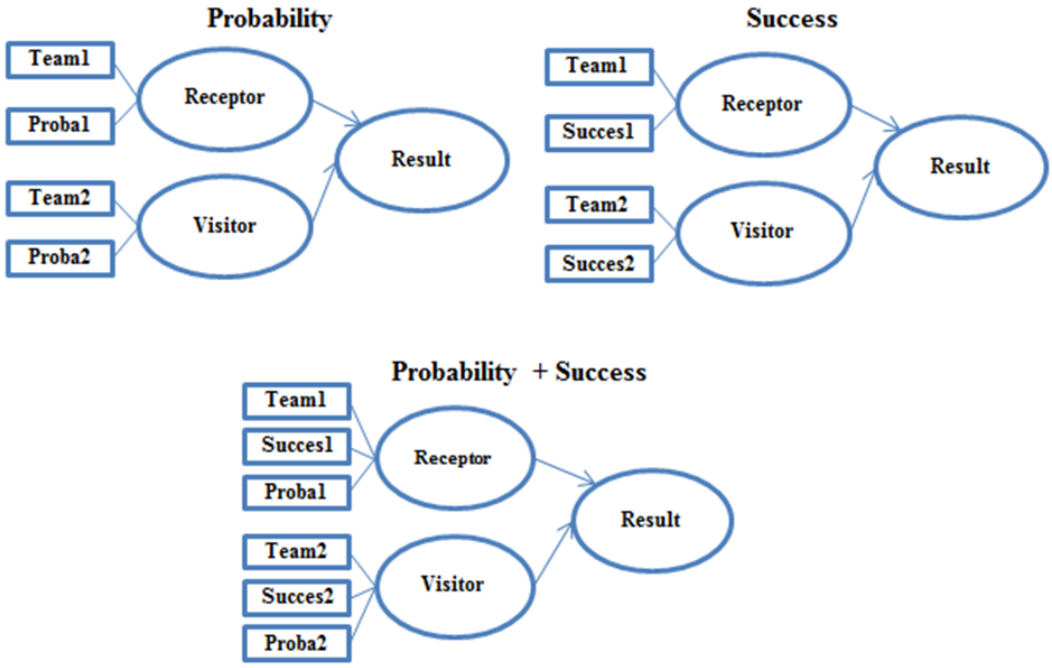

Fig. 8. Different models used

For that, we used in our model, in a first time, only variables containing the probability of winning for each team in such game. Then we repeated the same experiment by introducing variables containing this time the success of every team. Finally we mixed those models considering both variables. We applied those model on a collection composed of a variety of 240 games played this season in English premier league from the eleventh days, until the thirtieth one (table 5). We ignored the ten first days because the amount of data is not sufficient to compute success by using PLS-PM, so the results during those first journeys were not reliable to estimate the effectiveness of the model.

Table 5. Extract showing the kind of data we used in this experimentation

Team1	Suc1	Prob1	Team2	Suc2	Prob2	Day	Result
WestHam	-1.2342	1.94	Cardiff	-0.3484	4.02	11	1
WBA	-0.4313	2.29	Southampton	0.7968	3.12	11	-1
Swansea	-0.8592	4.37	ManUtd	0.7436	1.86	11	-1
Sunderland	-1.0079	2.21	Fulham	-0.4017	3.30	11	-1
Norwich	-0.3735	3.17	Everton	1.03976	2.31	11	0
Liverpool	1.4007	1.39	Stoke	-0.82890	8.67	11	1

Table 5 represents success at home (Suc1) and away (Suc2), we computed before, by applying the PLS-PM model and the probability of winning that match. The variable Result equals 1 when the receptor won, 0 when the match finished by a draw, and -1 when the visitor won. By validating our model using bootstrapping, we got the following results of model respectively based on probability, on success, and on both of them:

Table 6. Results of Bootstrapping

Model	Original	Mean.Boot	Std.Error	perc.025	perc.975
Probability	0.2155	0.2271	0.0456	0.1363	0.3158
PLS-PM	0.3384	0.3467	0.0465	0.2514	0.4171
Probability + PLS-PM	0.3010	0.3133	0.0463	0.2186	0.4011

As we can see in table 6, the lowest original value of R square obtained, depending on result, is when we used probability (21%), it means that the performance of the model based on probability is low. Poisson distribution depends on short term time periods and in this experiment, it loses its advantage, because we considered a long duration. Therefore, when considering the overall performance of a team over long periods, highest values of original R square rely on the success scores based on PLS-PM model.

4 Conclusion

Thanks to our different experimentations, we have shown that PLS-PM is a efficient method for prediction and prognostication. Starting by choosing the good Latent and manifest variables, then creating adequate relations between those variables, allowed us to get a model very close to the probabilistic one, which is considered to be the highest performance model in the state of the art. The main interest of PLS-PM is that we are able to introduce heterogeneous data in our model, and that is what permitted us to predict some draws by adding text data extracted from twitter. Draw case is very difficult, even the probabilistic model is not able to properly predict this kind of result. Our investigations showed that we cannot reach a good result without a lot of information. Nevertheless, such a result is known to be the weak point of PSL-PM, and we clearly observed it when we were not able to predict any match before the 10th journey. It means that, due to the lack of data, we could not predict 100 games. But even the probabilistic model has the same problem, despite that it does not need such an important mass of data as PLS-PM to predict games. There is a room of improvement in our results, and we intend to exceed the performance of probabilistic model as future work by trying to include more information that influences a football game to see how much game we can predict properly.

References

1. Henseler, J. *On the convergence of the partial least squares path modeling algorithm* Published online: 1 August 2009, DOI 10.1007/s00180-009-0164-x

2. Hu M. and Liu B. *Mining and Summarizing Customer Reviews.* Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA,
3. Khazaal Y., Chatton A., Billieux J., Bizzini L., Monney G., Fresard E., Thorens G., Bondolfi G., El-Guebaly N., Zullino D., Khan R. *Effects of expertise on football betting* Substance Abuse Treatment, Prevention, and Policy 2012 7:18.
4. Liu B., Hu M. and Cheng J. *Opinion Observer: Analyzing and Comparing Opinions on the Web.* Proceedings of the 14th International World Wide Web conference (WWW-2005), May 10-14, 2005, Chiba, Japan.
5. Maher M. J. *Statistica Neerlandica Modelling association football scores* Published online: 29 APR 2008, DOI: 10.1111/j.1467-9574.1982.tb00782.x
6. Sanchez G. *PLS path modeling with R*
7. Wold. S., Eriksson L., Trygg J., Kettaneh N. *The PLS method partial least squares projections to latent structures and its applications in industrial RDP (research, development, and production)* Submitted version, June 2004

Author Index

Aivar Sootla, 51

Abderrafiaa Koukam, 149
Alexandre Ravet, 61
Anderson Rocha, 159
Anett Seeland, 129
Anne Boyer, 1
Armelle Brun, 1

Bernard Manderick, 139
Bichen Shi, 71

Cynthia Rudin, 121

Fabio A. Faria, 159
Fabrice Lauri, 149
Frank Kirchner, 111, 129

Georgiana Ifrim, 71
Guy-Bart Stan, 51

Hendrik Wöhrle, 111
Hendrik Woehrle, 129

Jean-Valère Cossu, 91
Jiawei Zhu, 149
João Gama, 41
João Mendes-Moreira, 41
Johannes Teiwes, 129

Koen W. De Bock, 81
Kristof Coussement, 81

Luís Trigo, 169
Luis Moreira-Matias, 41

Madalina M. Drugan, 139
Marharyta Aleksandrova, 1
Mario Michael Krell, 111, 129
Markus Hegland, 101
Michel Ferreira, 41
Mohamed El Hamdaoui, 91

Neil Hurley, 71

Oleg Chertov, 1
Olivier Pietquin, 11

Pauli Miettinen, 31
Pavel Brazdil, 169

Ricardo da S. Torres, 159

Saba Q. Yahyaa, 139
Sanjar Karaev, 177
Saskia Metzler, 31
Simon Lacroix, 61
Sirko Straube, 111
Stijn Geuens, 81

Theja Tulabandhula, 121
Timothé Collet, 11

Valeriy Khakhutskyy, 101
Vincent Hilaire, 149

Wei Pan, 51

Ye Yuan, 51
Yury Kashnitsky, 21