ESTIMATION OF TEAM'S STRENGTH FOR HANDBALL GAMES PREDICTIONS

Florian Felice[1 2], Christophe Ley[1]

[1] Department of Mathematics, University of Luxembourg, (e-mail: florian.felice@uni.lu, christophe.ley@uni.lu)

[2] Work not related to Amazon

ABSTRACT: In this work, we present Statistically Enhanced Learning (SEL), a general feature extraction approach to improve any type of learning technique, whether it is statistical or machine learning. By adding highly informative covariates, which are obtained as statistical estimates rather than directly observed, SEL improves model learning for any type of data (tabular, computer vision, text). We will discuss the general idea and refer to existing feature extraction methods that actually can be shown to fall under the umbrella of SEL. In particular, we will see how SEL allows improved predictions of handball tournaments and discuss how it can be used to derive a metric for teams' or players' strengths.

KEYWORDS: Statistically Enhanced Learning, Feature extraction, Handball, Team's strength

1 Introduction

Statistically Enhanced Learning (Felice *et al.*, 2023) is a framework that aims to formalize the feature extraction step of the data processing in a machine learning project. Classified in different categories, SEL approaches can include proxy variables (Wooldridge, 2009) as well as statistical features that represent non-measurable quantities (Groll *et al.*, 2019). In particular, in sports predictions, factors such as the strength of the opposing teams are crucial elements but can not be measured objectively. Ley *et al.*, 2019 proposed a bivariate Poisson model to represent the outcome of football games. They estimate the location parameter λ for each team via Maximum Likelihood Estimation approach. They assume that one can derive the ability of the opposing teams using the formula $\lambda = \beta_0 + r_i - r_j$ where β_0 is a constant intercept, r_i and r_j are the abilities for the home and away teams. These parameters are later included in the training data set.

In the context of handball, Groll *et al.*, 2020 analyzed historical games to determine the best probability distribution to model the number of goals scored in handball matches. Given the level of under-disperson observed, they concluded that a Gaussian distribution with low variance is the most appropriate.

In the following, we will extend the work done by Groll *et al.*, 2020 and consider an additional discrete probability distribution. From this distribution, we will generate a metric representing the strength of a team. This metric can then be considered as a new SEL variable to be added in a training set to predict the outcome of handball games.

2 Modelling handball games with Conway-Maxwell-Poisson

As a fast-paced sport, handball can record a large number of goals during a 60 minute game (on average 27.9 for women and 29.8 for men). To model the number of goals scored, the traditional Poisson distribution assumes equi-distribution (i.e. $\mathbb{E}(X) = \mathbb{V}(X)$), however, historical data rarely satisfy this assumption.
Therefore, we compare here different distributions: the Gaussian and Negative Binomial distributions (as in Groll *et al.*, 2020) and the Conway-Maxwell-Poisson distribution (Sellers, 2023). The latter is a generalization of the common Poisson distribution, which can handle under- and over-dispersion.

Table 1: Comparison of log-likelihood evaluated on scored goals by Metz handball over season 2022/2023.

Distribution	Log-likelihood	AIC
Conway-Maxwell-Poisson	-127.66	259.31
Gaussian	-127.39	258.78
Negative Binomial	-127.36	258.72

As we can observe in Table 1, the three distributions seem to equivalently fit our data. However, we can notice the slight superiority of the Conway-Maxwell-Poisson distribution. Thus, given the continuous nature of the Gaussian distribution, we decide to discard it. Indeed, it can return non-integer values but, more problematically, it is defined on the real line which includes negative values. Furthermore, in our experiments, the Conway-Maxwell-Poisson distribution consistently showed superiority over the Negative Binomial.
As a result, and considering its flexibility to handle any of the under-, over-, and equi-dispersion, we consider the Conway-Maxwell-Poisson distribution for the rest of this work.

3 Estimation of team's strength

Adopting the selected Conway-Maxwell-Poisson (CMP) distribution, we use its parameters to represent the strength of a team both for attack and defense. With the distribution of scored goals following a CMP distribution, $Y_a \sim CMP(\lambda_a, \nu_a)$, the parameter $\lambda_a > 0$ can act as a location parameter and $\nu_a \geq 0$ as the dispersion parameter. We define the attack strength of a team as:

$$s_a = \frac{\log(\lambda_a)}{\nu_a}. \tag{1}$$

We want to penalize for irregular performances, hence we use ν_a as the denominator so the higher the irregularities the lower the attack strength score.
Similar to attack, the distribution of goals conceded by a team follows a CMP distribution, $Y_d \sim CMP(\lambda_d, \nu_d)$. However, the strength of a team's defense is inversely proportional to the goals it concedes. Thus, we define the defense strength s_d as:

$$s_d = \frac{\nu_d}{\log(\lambda_d)}. \tag{2}$$

A team is considered strong when it can perform well in attack and defense. We can consider the overall strength of a team as the product of attack and defense strengths, formally:

$$s = s_a \cdot s_d = \frac{\log(\lambda_a) \cdot \nu_d}{\nu_a \cdot \log(\lambda_d)}. \tag{3}$$

Empirically, we illustrate these results in Table 2 with European female clubs. We can observe that the teams considered the strongest are the leading clubs in their country and strong competitors in the European Champions League.

4 Conclusion

Using the parameters from the fitted Conway-Maxwell-Poisson distribution, we can estimate the strength of a team. The estimated parameters constitute new SEL variables to be added to the training set for match predictions (Felice, 2023). These estimations can also be applied, with a similar logic, to player's performance and derive the player's strength. Our results and conclusions are derived from men and women club's data, but we note that they also apply to national teams for international competitions.

Table 2: Top 10 strongest female teams in Europe for season 2022/2023.

Team	Avg. scored	Avg. conceded	Attack strength	Defense strength	Strength
Győri Audi ETO KC	33.32	24.32	3.49	3.16	11.00
Vipers Kristiansand	37.62	26.38	3.57	3.07	10.96
Podravka Vegeta	30.50	21.75	3.39	3.21	10.89
Metz handball	33.58	24.00	3.47	3.12	10.85
Team Esbjerg	33.33	24.67	3.48	3.11	10.83
SG BBM Bietigheim	34.63	25.21	3.54	3.05	10.80
HC Dunajskà Streda	29.37	22.62	3.38	3.15	10.63
Herning-Ikast Håndbold	28.71	23.29	3.39	3.13	10.61
DVSC Schaeffler	30.83	24.11	3.39	3.12	10.59
CSM București	33.13	25.83	3.48	3.05	10.58

References

FELICE, FLORIAN. 2023. Prediction of Handball Games with Statistically Enhanced learning via Estimated Teams' Strength. *arXiv preprint*.

FELICE, FLORIAN, LEY, CHRISTOPHE, GROLL, ANDREAS, & BORDAS, STEPHANE. 2023. Statistically Enhanced Learning - a Formalization Framework of Feature Extraction Techniques. *arXiv preprint*.

GROLL, ANDREAS, LEY, CHRISTOPHE, SCHAUBERGER, GUNTHER, & VAN EETVELDE, HANS. 2019. A hybrid random forest to predict soccer matches in international tournaments. *Journal of Quantitative Analysis in Sports*, **15**(4), 271–287.

GROLL, ANDREAS, HEINER, JONAS, SCHAUBERGER, GUNTHER, & UHRMEISTER, JÖRN. 2020. Prediction of the 2019 IHF World Men's Handball Championship–A sparse Gaussian approximation model. *Journal of Sports Analytics*, **6**(3), 187–197.

LEY, CHRISTOPHE, VAN DE WIELE, TOM, & VAN EETVELDE, HANS. 2019. Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches. *Statistical Modelling*, **19**(1), 55–73.

SELLERS, KIMBERLY F. 2023. *The Conway-Maxwell-Poisson Distribution*. Cambridge University Press.

WOOLDRIDGE, JEFFREY M. 2009. On estimating firm-level production functions using proxy variables to control for unobservables. *Economics Letters*, **104**(3), 112–114.