

Optimizing the Prediction Process: From Statistical Concepts to the Case Study of Soccer

Optymalizacja procesu przewidywania: od koncepcji statystycznych do studium przypadku piłki nożnej

Julia Skrzypczak

1 *Introduction*

2 *Statistical background*

- Regression
- Prediction for individual subjects/teams
- Application to the Case of Soccer Prediction

3 *Results*

4 *Bibliography*

Dokonamy przewidywań na podstawie danych panelowych, które zawierają informacje o różnych obiektach wchodzących ze sobą w interakcje w dwóch przedziałach czasowych. Dla uproszczenia założymy, że badane obiekty nie zmieniają swoich właściwości pomiędzy ostatnim pomiarem a okresem przewidywań.

Introduction

Rozważymy dwa przedziały czasowe, w których będziemy dokonywać pomiary na dwóch niezależnych zmiennych X i Y .

season			
	previous	present	
observable	$X = \Delta C_X$	$Y = \Delta C_Y$	$Z = \Delta G_Z$
team strength	S_X	S_Y	S_Z
# matches	$N_X (34)$	$N_Y (\leq 17)$	$N_Z (17)$

 prediction

Zacniemy od liniowej zależności między niezależną zmienną Y i zależną zmienną Z :

$$Z = bY$$

Zakładamy, że pierwszy moment wszystkich zmiennych jest równy 0.

Chcemy zminimalizować $\langle (Z - \hat{Z})^2 \rangle$ w zależności od b , gdzie $\hat{Z} = bY$ jest przewidywaniem Z .

$$b_{opt} = \text{corr}(Y, Z) \sqrt{\frac{\text{Var}(Z)}{\text{Var}(Y)}}$$

Po podstawieniu b_{opt} otrzymujemy optymalną kwadratową wariancję oznaczoną $\chi^2(Y)$:

$$\chi^2(Y) = \text{Var}(Z)[1 - [\text{corr}(Y, Z)]^2],$$

oraz współczynnik korelacji:

$$\text{corr}(Y, Z) = \frac{\langle YZ \rangle}{\sqrt{\text{Var}(Y)\text{Var}(Z)}}.$$

Teraz mamy do czynienia z regresją dwuwymiarową

$$\hat{Z} = aX + bY.$$

Otrzymujemy

$$\chi^2(X, Y) = \chi^2(Y)[1 - [\text{corr}(X - Y, Z - Y)]^2]$$

oraz częściowy współczynnik korelacji

$$\text{corr}(X - Y, Z - Y) = \frac{\text{corr}(X, Z) - \text{corr}(X, Y)\text{corr}(Y, Z)}{\sqrt{1 - \text{corr}(X, Y)^2}\sqrt{1 - \text{corr}(Y, Z)^2}}.$$

Zmienne X, Y, Z oznaczają osiągnięcia zespołów podczas trzech przedziałów czasowych. Dla pierwszego z nich osiągnięcie i -tego zespołu oznaczmy jako x_i . Na tę wartość składa się siła tego zespołu $s_{X,i}$ oraz losowe okoliczności, których nie przewidzimy $\varepsilon_{X,i}$

$$x_i = s_{X,i} + \varepsilon_{X,i},$$

$$\text{Var}(X) = \text{Var}(S_X) + \text{Var}(\varepsilon_X).$$

Analogiczne zależności są dla $\text{Var}(Y)$, $\text{Var}(Z)$.

Dalej potrzeba obliczyć współczynniki korelacji, takie jak:

$$\text{corr}(Y, Z) = \frac{\text{corr}(S_Y, S_Z)}{\sqrt{1 + \text{Var}(\varepsilon_Y / \text{Var}(S_Y))} \sqrt{1 + \text{Var}(\varepsilon_Z / \text{Var}(S_Z))}}$$

Analogicznie dla $\text{corr}(X, Z)$, $\text{corr}(X, Y)$.

Siła zespołu może się zmienić między przedziałami czasowymi, więc $\text{corr}(S_Y, S_Z) < 1$.

Losowe zdarzenia wpływają na Y, Z , więc $\text{Var}(\varepsilon_Y), \text{Var}(\varepsilon_Z) > 0$.

Reszta rozważań bazuje na następującej tożsamości:

$$\frac{\text{corr}(X, Y)}{\text{corr}(X, Z)\text{corr}(Y, Z)} = \frac{\text{corr}(S_X, S_Y)}{\text{corr}(S_X, S_Z)\text{corr}(S_Y, S_Z)} \left(1 + \frac{\text{Var}(\varepsilon_Z)}{\text{Var}(S_Z)} \right).$$

Dla uproszczenia estymujemy S_Z zamiast Z . Wtedy mamy $\tilde{\chi}^2(X, Y)$.
Zakładamy, że $\text{Var}(\varepsilon_Z) = 0$.
Otrzymujemy równanie

$$\chi^2(X, Y) = \tilde{\chi}^2(X, Y) + \text{Var}(\varepsilon_Z).$$

Zakładamy, że w trakcie sezonu siła zespołu się nie zmienia $S_Y \approx S_Z$. Stąd

$$\text{corr}(S_X, S_Y) = \text{corr}(S_X, S_Z).$$

Zatem

$$\tilde{\chi}^2(X, Y) = \text{Var}(S_Z) \frac{(1 - [\text{corr}(Y, S_Z)]^2)(1 - [\text{corr}(X, S_Z)]^2)}{1 - [\text{corr}(X, Y)]^2}.$$

Ta zależność będzie potrzebna nam do szacowania jakości przewidywań. Mimo dokonanych założeń, jest ona kluczowa.

W późniejszym przykładzie będziemy korzystać z danych zawierających wszystkie mecze Bundesligi w sezonach 1995/96 oraz 2010/11. Podczas sezonu każda drużyna rozgrywa 34 mecze.

Naszym celem będzie przewidzieć wynik z_i dla każdej drużyny w drugiej połowie sezonu na podstawie N_Y wyników meczy y_i z pierwszej połowy sezonu oraz $N_X = 34$ wyników x_i z poprzedniego sezonu.

Gdzie z_i to różnica goli drużyny i po N_Z meczach znormalizowana przez N_Z .

Chcemy przewidzieć różnicę goli Z (alternatywnie siłę zespołu S_Z).
Za zmienne X, Y przyjęliśmy różnicę goli w odpowiednich przedziałach czasowych.
Niech ΔG oznacza różnicę goli.

Podczas dalszych rozważań ograniczymy się do jednowymiarowego problemu $Y \rightarrow Z$.

Niech $N_Y = N_Z = 17$

Wyznacznikiem jakości naszych przewidywań będzie $\text{corr}(Y, Z)$.

Z naszych danych mamy:

$$\text{corr}(Y = \Delta G_Y, Z = \Delta G_Z) = 0.56$$

Czy możemy zwiększyć współczynnik korelacji?

Niech:

$C_{\pm}$ oznacza szanse na zdobycie goli,

$G_{\pm}$ oznacza liczbę goli,

$p_{\pm}$ oznacza współczynnik efektywności zdobywania goli,

gdzie $+$ wskazuje na rozważaną drużynę a $-$ na drużynę przeciwną.

Wtedy

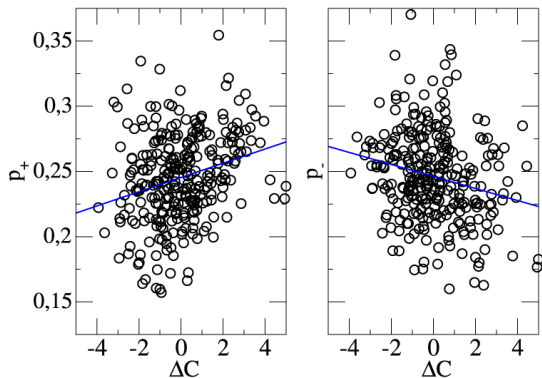
$$G_{\pm} = C_{\pm} \cdot p_{\pm}$$

Stąd $p_+ = \frac{G_+}{C_+}$ oznacza prawdopodobieństwo tego, że drużyna wykorzysta szansę na zdobycie gola i przekształci ją w prawdziwego gola.

Wówczas $1 - p_-$ oznacza prawdopodobieństwo tego, że drużynie uda się nie stracić gola, kiedy przeciwnicy będą mieli szansę na zdobycie bramki.

$$\langle p_{\pm} \rangle = 0,24$$

Application to the Case of Soccer Prediction

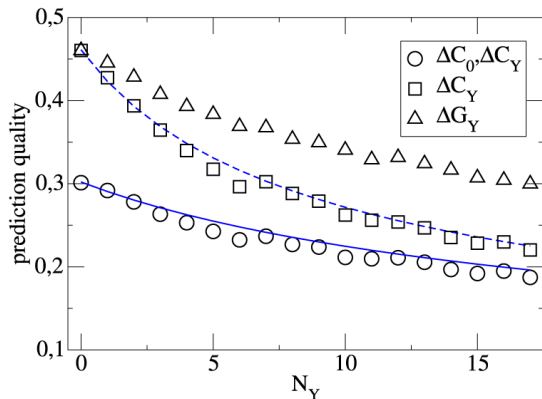


Jeżeli zdefiniujemy ΔC jako $\Delta C = C_+ - C_-$, otrzymamy

$$\text{corr}(Y = \Delta C, Z = \Delta G_Z) = 0.65$$

Results

The prediction quality of the team strength determined via $\sqrt{\tilde{\chi}^2(X, Y)}$ as a function of the number of match days N_Y



Na podstawie konkretnych danych z Bundesligi chcemy przewidzieć różnicę goli w drugiej połowie sezonu.

$$\Delta \tilde{G}_Z = a(N_Y) \Delta C_X + b(N_Y) \Delta C_Y(N_Y)$$

Mamy

$$\begin{aligned} a(N_Y = 0) &= \frac{17 \text{corr}(\Delta C_X, \Delta G_Z)}{\text{Var}(\Delta C_X)} = \\ &= \frac{17 \text{corr}(S_X, S_Z) \sqrt{\text{Var}(S_X) \text{Var}(S_Z)}}{\text{Var}(\Delta C_X)} \approx 3,8 \end{aligned}$$

N_Y	$a(N_Y)$	$b(N_Y)$
0	3.71	0
4	3.20	0.82
8	2.60	1.70
12	2.23	2.30
17	1.86	2.77

doi:10.1371/journal.pone.0104647.t002

Results

	$17\Delta G_Y$	$17\Delta G_Z$	$17\Delta G_{Z,est}(N_Y = 0)$	$17\Delta G_{Z,est}(N_Y = 17)$	$17\Delta G_{Z,est}(N_Y = 17)$ plus market value
B. München	23	24	10	21	23
Bremen	18	12	11	15	14
Hamburg	11	10	3	9	10
Leverkusen	16	1	2	9	8
Schalke	9	14	8	12	11
Karlsruhe	-2	-13	-8	-6	-7
Hannover	-1	-1	3	-1	-2
Stuttgart	-1	1	9	5	6
Frankfurt	-4	-3	2	-3	-4
Dortmund	-4	-8	0	0	2
Wolfsburg	0	12	-4	-5	-2
Hertha	-5	0	-5	-8	-5
Bochum	-2	-4	-1	-4	-7
Bielefeld	-19	-6	-6	-11	-10
Rostock	-10	-12	-8	-11	-13
Nürnberg	-7	-9	1	1	1
Cottbus	-10	-11	-8	-10	-12
Duisburg	-12	-7	-8	-13	-12

The estimation in the final column also involves information about the market value. The actual goal differences of the first half of that season and the second half are included in the first two columns, respectively.

doi:10.1371/journal.pone.0104647.t003

Jeżeli ΔG jest utożsamiana z siłą drużyny, to możemy użyć jej do estymacji indywidualnych meczy.

Zobaczmy teraz przykład na konkretnym meczu z naszych danych:
18 dzień meczowy sezonu 2007/08 Cottbus vs Leverkusen

$$(-12)/17 - (+8)/17 + 0.3 = -0.9$$

Prawdziwy wynik tego meczu 2:3

Optimizing the Prediction Process: From Statistical Concepts to the Case Study of Soccer - Andreas Heuer ,Oliver Rubner