Poisson-Gamma model

Leo Bastos (Fiocruz) & Joel Rosa (Rockfeller University)

24-07-2014

Summary

- Brief review
- Prediction model based on number of goals
  - Building the model
  - Priors
- Strength factor
  - Fuzzy C-means clustering
- Results
  - 2010 FIFA World Cup
- Discussion

Brief review

This is not an extensive literature review.

- Dixon and Coles (1997) - Independent Poisson (English league)
  - Crowder, Dixon, et al. (2002) - Dynamic model
  - Rue and Salvense (2000) - Generalized DLM
- Dyte and Clarke (2000) - log-linear model at (1998 FIFA)
  - Suzuki et al. (2009) - Fully subjective Poisson-Gamma (2006 FIFA)
  - Bastos and Rosa (2013) - Poisson-Gamma with Fuzzy clustering (2010 FIFA)
- Brillinger (2008) - Trinomial model (Brasileirão 2006)
- Ramirez-Hassan and Cardona-Jimenez (2014) - Categorical-dirichlet model (2014 FIFA)

The number of goals of each team

- Suppose team A will play against team B.
- Let N_A be the number of goals of team A.
- Let N_B be the number of goals of team B.
- The winner is the one with more goals, if $N_A = N_B$ it is a draw.
- Modelling $p(N_A, N_B)$ allow us to infer the macth outcome.

Probabilities of the possible outcomes

$$
\begin{aligned}
P(\text{“Team A wins”}) &= P(N_A > N_B) \\
&= \sum_{a=1}^{\infty} \sum_{b=0}^{a-1} P(N_A = a, N_B = b), \\
P(\text{“Draw”}) &= P(N_A = N_B) \\
&= \sum_{z=0}^{\infty} P(N_A = z, N_B = z), \\
P(\text{“Team B wins”}) &= \sum_{b=1}^{\infty} \sum_{a=0}^{b-1} P(N_A = a, N_B = b).
\end{aligned}
$$

Building the model

- Joint distribution

$$p(n_A, n_B) = \int \int p(n_a, n_b|\lambda_a, \lambda_b) p(\lambda_a, \lambda_b) d\lambda_a d\lambda_b.$$

  where λ_K is a scoring rate of team K.
- Conditional independence (too strong? maybe!)

$$p(n_a, n_b|\lambda_a, \lambda_b) = p(n_a|\lambda_a, \lambda_b) p(n_b|\lambda_a, \lambda_b)$$

- Poisson distribution

$$N_k|\lambda_k \sim Pois(\lambda_k)$$

  (Ignoring the scoring rate of the oponent)
- The problem reduces to choosing a good prior for the scoring rates.

Gamma prior

- We assume a Gamma prior for the scoring rate for each team

$$\lambda_k \sim \mathit{Gamma}(\alpha_k, \beta_k), \qquad \forall k$$

- The predictive distribution for the number of goals is

$$N_K \sim \mathit{NegBin}\left(\alpha_k, \frac{\beta_k}{\beta_k + 1}\right)$$

where $N_K \perp N_{K'}$.

Example

- Suppose team A will play against team B

Team	Sample mean	Standard deviation
Team A	3.0	1.0
Team B	1.0	1.0

- Then we can calculate
- $\alpha_a = 9.0$ and $\beta_a = 3.0$
- $\alpha_b = 1.0$ and $\beta_b = 1.0$.

Probabilities

Team A / Team B	0	1	2	3	4
0	0.0375	0.0188	0.0094	0.0047	0.0023
1	0.0845	0.0422	0.0211	0.0106	0.0053
2	0.1056	0.0528	0.0264	0.0132	0.0066
3	0.0968	0.0484	0.0242	0.0121	0.0060
4	0.0726	0.0363	0.0181	0.0091	0.0045
5	0.0472	0.0236	0.0118	0.0059	0.0029
6	0.0275	0.0138	0.0069	0.0034	0.0017
7	0.0147	0.0074	0.0037	0.0018	0.0009

Probabilities

- The most likely outcome is Team A 2 x 0 Team B
- The probability of team A win the game is 0.7503
- The probability of team B win the game is 0.1249
- The probability of a draw is 0.1248

Dynamic modelling

- Dynamic model can also be fitted

$$p(n_k^{(t)}|D_{t-1}) = \int_0^\infty p(n_k^{(t)}|\lambda_k)p(\lambda_k|D_{t-1})d\lambda_k$$

- Poisson-Gamma model leads to

$$N_K^{(t)}|D_{t-1} \sim NegBin\left(\alpha_k^{(t-1)}, \frac{\beta_k^{(t-1)}}{\beta_k^{(t-1)}+1}\right).$$

where

- $\alpha_k^{(t-1)} = \alpha_k^{(0)} + \sum_{i=1}^{t-1} n_k^{(i)}$
- $\beta_k^{(t-1)} = \beta_k^{(0)} + (t-1)$

Eliciting priors for scoring rates

- Instead of elicit (α, β), we elicit (m, v)
- m_k is an average scoring rate of team K, and v_k its variance.
  - $m_k^{(0)} = \bar{y}\delta_k, \qquad \delta_k \in (0, 1)$
  - $v_k^{(0)} = S_y^2$
- $\bar{y}$ and S_y^2 came from the qualyfies for the World Cup.
- δ_k is a strength factor.

Strength factor

Variable	Description
X_1	Total of points obtained in World Cups;
X_2	Lowest rank position obtained since the creation of FIFA ranki
X_3	Range between lowest and highest rank;
X_4	Number of points in FIFA ranking in June of 2010;
X_5	Number of players that participated in 2009-10 Champions Lea
X_6	Performance in World Cup Qualifying games (% of points).

Strength factor (Fuzzy C-means cluster)

- We decided to use elements of fuzzy set theory to discriminate national teams.
- Fuzzy clusters were built using the Fuzzy C-means algorithm.
  - $g \in 1, 2, \ldots, G$ clusters
  - $u_g(\mathbf{x}_k)$ membership degree of team k $(\sum_g u_g(\mathbf{x}_k) = 1)$
- The algorithm is based on a c-means objective function

$$J_m = \sum_{k=1}^{K} \sum_{g=1}^{G} [u_g(\mathbf{x}_k)]^m \, ||x_k - \mathbf{v}_g||^2$$

  - $m \geq 1$ is a degree of fuzziness
  - $\mathbf{v}_g$ is the centroid of each cluster

Strength factor (Fuzzy C-means cluster)

- Using literature review and cluster stability measures
- The degree of fuzziness m was set to 1.25.
- The number of clusters G was set to 4.
- We use the R package *fanny*
- The strength factor was then defined by

$$\delta_k = \sum_{j=1}^{4} w_j u_j(\mathbf{x}_k)$$

  where w_j is the average FIFA ranking for each cluster.
- The strength factors are then normalized to be in (0,1).

National Fuzzy clusters

Clusters	Teams
1	South Africa, Algeria, Korea DPR, Slovakia, Slovenia, Hondura
2	Germany, Spain, Holland, England, Italy.
3	Argentina, Brazil, France, Portugal.
4	Australia, Cameron, Chile, Korea Republic, Ivory Coast, Denma

World cup predictions

Team A (Probability)	Actual score	Team B (Probability)
South Africa (0.273)	1 x 1	Mexico (0.455)
Brazil (0.576)	2 x 1	Korea DPR (0.204)
Brazil (0.504)	3 x 1	Ivory Coast (0.275)
Portugal (0.393)	0 x 0	Brazil (0.3975)
Brazil (0.523)	3 x 0	Chile (0.255)
Holland (0.203)	2 x 1	Brazil (0.5769)
Holland (0.204)	0 x 1	Spain (0.575)

DeFinetti measure

The De Finetti measure is the Euclidean distance between the indicator vector of the actual score, e.g. $(1, 0, 0)$, and the probability vector, e.g. $(0.333, 0.334, 0.333)$.

2010 WC	Static	Dynamic
Group stage	0.6258	0.6323
Knockout stage	0.5668	0.5768
Overall	0.6110	0.6185

A full naive prediction $= 0.6667$

Comments

- Football predicton is hard, the uncertainty is HUGE!
- But, it is a nice playground for model development.
- Drawbacks in the methodology we used:
  - Goals in a game by each team are not independent
  - Draws are under estimated
  - No expert opnions during the tournament (although it is possible)
  - Probably several others
- Luckly we didn't predict the results for the 2014 World Cup.

Future paths to explore

- Fuzzy clustering using time-depend variables.
- Strenght factor depending on the opponent
- Bivariate correlated Poisson (In progress)

$$N_A = Z_A + Z_D$$

$$N_B = Z_B + Z_D$$

$$Z_i \sim Pois(\lambda_i), \quad Z_i \perp Z_j$$

$$\lambda_i \sim Gamma(a_i, b_i)$$

- Add expert priors (Bayesian melding like)
  - Poole and Raftery (2000)

Thank you

- This presentation, the data set and scripts are (will be) freely available in my homepage
- Leo Bastos: lsbastos@fiocruz.br
- My homepage: `http://www.procc.fiocruz.br/Members/lsbastos`