

Modelo de fuerza deportiva

NOTA TÉCNICA

Contenido

1	Motivación	3
2	El modelo de fuerza deportiva	3
2.1	Supuestos del modelo	3
2.2	Críticas al modelo	4
2.3	Formulación	4
2.4	Estimación de los parámetros del modelo	5
3	Predicción de resultados	6
4	Cálculo de probabilidades	6
4.1	Probabilidad de la diferencia de goles	6
4.2	Probabilidad del signo de un partido	6
4.3	Probabilidad del resultado exacto de un partido	7
4.4	Probabilidad de una competición	7
4.5	Ejemplos numéricos	7

Copyright	2010, Bayes Inference S.A.
Título	Modelo de fuerza deportiva
Categoría	NOTA TÉCNICA
Edición	2010/06/07 18:04:00
Claves	fútbol, equipo, partido, resultado, gol, modelo, parámetro, fuerza, probabilidad, factor campo, varianza, esperanza, previsión
Distribución	Pública

La presente Nota Técnica ha sido creada por **Bayes Inference, S.A.**, en adelante **Bayes**. En consecuencia, su contenido y diseño es de la exclusiva propiedad de **Bayes**, correspondiéndole los derechos de titularidad que se derivan de la misma.

1 Motivación

La necesidad de representar el fútbol mediante un modelo matemático surgió en la compañía en los años 90 ante el escaso grado de acierto de la modelación tradicional de los resultados deportivos, basada en variables *dummies*, a la hora de explicar y prever comportamientos sociales como las ventas de un diario deportivo, las llamadas a un *call-center*, los pedidos de comida rápida, etc.

Pronto se observó que las variables que recogían la victoria, el empate o la derrota de un equipo, aun ponderadas por el resultado, presentaban serias deficiencias, ya que no tenían en cuenta:

- ✓ El nivel deportivo del rival.
- ✓ El estado de la competición.
- ✓ La evolución temporal de los equipos implicados.

Lo anterior motivó el desarrollo de un modelo que diera solución a los problemas planteados y ayudara a mejorar cómo influyen los resultados deportivos en multitud de aspectos del comportamiento humano.

2 El modelo de fuerza deportiva

La fuerza deportiva de un equipo en un momento dado representa la capacidad que tiene de marcar goles. Por tanto, el resultado de un partido depende de las fuerzas de los equipos que se enfrentan.

Todo el desarrollo que se expone a continuación está aplicado y particularizado al fútbol, deporte predominante en la mayor parte de los países de Europa y Sudamérica y, particularmente, en España. No obstante, no resultaría complicado adaptarlo a otros deportes como el baloncesto, el tenis, el golf, etc.

2.1 Supuestos del modelo

El modelo propuesto se basa en los siguientes supuestos:

- ✓ Sólo tiene en cuenta los resultados deportivos obtenidos por los equipos en el pasado.
- ✓ La fuerza de un equipo es única para todas las competiciones y fases en las que participa.
- ✓ La fuerza es dinámica y adaptativa, es decir depende de los resultados pasados, dando más peso a los resultados más recientes.
- ✓ El sistema de fuerzas de un conjunto de equipos es un juego de suma cero.
- ✓ La fuerza de un equipo tiene métrica: representa la diferencia de goles esperada para un partido jugado en campo neutral entre dicho equipo y un rival de fuerza nula.
- ✓ El modelo tiene dos parámetros generales, que representan la ventaja que otorga el factor campo y la innovación que el resultado de un partido introduce en la fuerza de los equipos enfrentados. Ambos son constantes.
- ✓ El resultado de un partido modifica la fuerza de los equipos enfrentados de forma lineal; es decir, si la diferencia de goles se duplica, la variación de la fuerza se duplica igualmente.

2.2 Críticas al modelo

Los anteriores supuestos constituyen un conjunto de limitaciones o restricciones que conllevan las siguientes críticas al modelo:

- ✓ No tiene en cuenta otros factores objetivos que pueden influir en la fuerza deportiva, tales como: los jugadores que participan en cada partido, el presupuesto anual del club, el número de aficionados que acuden a ver el partido, la meteorología, etc.
- ✓ Al ser constante por equipo, la fuerza no permite recoger el diferente estado anímico que puede tener un equipo en las diferentes competiciones en las que participa, o la diferente intensidad con la que puede jugar en función de la fase en que se encuentre dentro de una competición.
- ✓ La suma de las fuerzas de todos los equipos es igual en todos los momentos temporales, ya que es un juego de suma cero; esto impide la comparación del nivel colectivo entre dos épocas diferentes. Sin embargo, sí permite realizar comparaciones geográficas para determinar, por ejemplo, si la liga española es más fuerte que la holandesa, o viceversa.
- ✓ No permite establecer un factor campo diferente por equipo o, más concretamente, por estadio, de manera que todos los equipos tienen la misma ventaja por jugar en casa. Además, este parámetro es constante en el tiempo, no evoluciona.
- ✓ El parámetro de innovación del modelo es constante, por lo que no depende del tiempo transcurrido desde el último partido que jugó un equipo. Parece lógico suponer que cuanto mayor sea este tiempo, más innovación introducirá el resultado en las fuerzas, ya que puede dar lugar a un mayor número de eventualidades: una lesión, un fichaje, etc.
- ✓ No tiene en cuenta factores subjetivos como que pueden influir en el resultado, como el estado anímico de los jugadores (en función, por ejemplo, de lo que haya en juego en el partido en cuestión), el cansancio de los jugadores (que dependerá del número de partidos que hayan jugado en la temporada o del tiempo de descanso desde el último partido jugado), etc.

2.3 Formulación

Sean F_t^a , F_t^b y c , respectivamente, las fuerzas deportivas de los equipos a y b en el momento t , y el factor campo. Sea, además, $G_t^{a,b}$ el resultado del enfrentamiento de los equipos a y b en el momento t y en el campo de a . La formulación del modelo es:

$$G_t^{a,b} = F_t^a - F_t^b + c + e_t \quad [1],$$

$$e_t \sim N(0, \sigma^2)$$

$$Def.: G_t^a = G_t^{a,b} - \mu_t = F_t^a + e_t \quad [2]$$

$$\nabla G_t^a = (1 - \theta B)e_t \quad [3]$$

La ecuación [1] describe el resultado, visto como diferencia de goles, entre un equipo a que actúa como local y un equipo b que actúa como visitante como la diferencia de sus respectivas fuerzas deportivas a las que se les suma el factor campo, más un término de error que sigue una distribución normal de media cero y varianza σ^2 . Partidos jugados en campo neutral tienen, naturalmente, un factor campo nulo.

La ecuación [2] define la diferencia de goles desde la perspectiva de cierto equipo a , es decir corregida por el factor campo y por la fuerza del rival. Éste puede actuar como local o como visitante. De ahí que la variable μ_t se define como:

$$\mu_t = \left\{ \begin{array}{ll} c - F_t^b, & \text{si } L(a) = 1 \\ -(c + F_t^b), & \text{si } (L(a) = 0 \wedge L(b) = 1) \\ -F_t^b, & \text{si } L(a) + L(b) = 0 \end{array} \right\} \quad [4]$$

donde $L(i)=1$ indica que i es local y $L(i)=0$ indica que i no es equipo local.

Diferenciando [2] tenemos que:

$$\nabla G_t^a = \nabla F_t^a + \nabla e_t \quad [5],$$

Si igualamos las partes derechas de [3] y [5] tenemos que:

$$\nabla F_t^a = (1 - \theta)e_{t-1} \quad [6].$$

O, lo que es lo mismo:

$$F_t^a = F_{t-1}^a + (1 - \theta)e_{t-1} = F_{t-1}^a + (1 - \theta)(G_{t-1}^{a,b} - E[G_{t-1}^{a,b}]) \quad [7],$$

donde el resultado esperado se define como:

$$E[G_t^{a,b}] = G_t^{a,b} - e_t = F_t^a - F_t^b + c \quad [8],$$

La interpretación de [7] es muy intuitiva: la fuerza de un equipo después de jugar un partido es igual a su fuerza antes de jugar más un factor de innovación resultante de multiplicar el complementario del parámetro de memoria por el error que comete el modelo en la previsión del resultado.

Utilizando [2], [7] puede reescribirse de forma alternativa como:

$$F_t^a = (1 - \theta)G_{t-1}^a + \theta F_{t-1}^a \quad [9].$$

La interpretación de [9] es igual de sencilla: la fuerza de un equipo después de jugar un partido es igual a su fuerza antes de jugar, en una proporción θ , más el resultado corregido del factor campo y de la fuerza de su rival, en la proporción complementaria $(1 - \theta)$.

2.4 Estimación de los parámetros del modelo

El modelo definido por [1], [2] y [3] consta de cuatro grupos de parámetros que necesitan ser estimados:

- ✓ La fuerza inicial de cada equipo F_0^i , es decir la fuerza que tenían en el primer partido de la muestra. Uno de los supuestos del modelo establece que, al igual que en el resto de instantes temporales, la suma de todas las fuerzas iniciales debe ser igual a cero: $\sum_i F_0^i = 0$.
- ✓ El parámetro de memoria, θ .
- ✓ El factor campo, c .
- ✓ La desviación típica de los residuos, σ .

La muestra utilizada para estimar los parámetros anteriores está formada por los partidos oficiales jugados desde 1992 correspondientes a las siguientes competiciones:

- ✓ Clubes: las principales ligas, copas y principales competiciones internacionales europeas.

- ✓ Selecciones: la principal competición de cada continente –Eurocopa, Copa América, Copa Asiática, Copa de África, Copa de Oro de la CONCACAF y Copa de Oceanía–, los mundiales y las fases de clasificación de dichos torneos.

En total, más de 100.000 partidos han sido utilizados en la estimación de los parámetros, dando como resultado los siguientes valores:

- ✓ El parámetro de memoria θ es de 0,98.
- ✓ El factor campo c es de 0,5. En los mundiales se aplica un factor campo diferente al anfitrión, juegue como local o como visitante, igual a 1,12 goles.
- ✓ La desviación típica σ de los residuos es de 1,78.

3 Predicción de resultados

El modelo propuesto posee capacidad predictiva. La distribución de probabilidad de la previsión del resultado de un partido es una normal cuya media es el resultado esperado y cuya varianza coincide con la del término de error: σ^2 :

$$G_t^{a,b} \sim N(E[G_t^{a,b}], \sigma^2)$$

$$E[G_t^{a,b}] = F_t^a - F_t^b + c$$

Así, la previsión del resultado de un partido entre dos equipos a y b es una variable aleatoria que sigue una distribución normal de media $E[G_t^{a,b}]$ y varianza σ^2 .

4 Cálculo de probabilidades

A partir de la distribución de probabilidad de la previsión de la diferencia de goles de un partido, es posible obtener las probabilidades de los distintos sucesos que pueden darse:

- ✓ La probabilidad de una diferencia de goles determinada.
- ✓ La probabilidad del signo de un partido: 1, X, 2.
- ✓ La probabilidad de un determinado resultado exacto.
- ✓ La probabilidad que cada uno de los equipos de una competición tiene de quedar en una determinada posición, pasar a la siguiente fase, etc.

4.1 Probabilidad de la diferencia de goles

La distribución normal es continua, mientras que el resultado de un partido es una variable discreta. Por ello, para obtener la probabilidad de que se dé una determinada diferencia de goles d en un partido, basta con calcular el área que queda entre dicha diferencia de goles y un rango de $\pm 0,5$ goles:

$$prob(G_t^{a,b} = d) = N(d + 0,5; E[G_t^{a,b}]; \sigma) - N(d - 0,5; E[G_t^{a,b}]; \sigma)$$

4.2 Probabilidad del signo de un partido

Del mismo modo, la probabilidad de cada uno de los signos posibles en un partido (1, X, 2) viene dada por:

$$prob(signo = 1) = 1 - N(0,5; E[G_t^{a,b}]; \sigma)$$

$$prob(signo = X) = N(0,5; E[G_t^{a,b}]; \sigma) - N(-0,5; E[G_t^{a,b}]; \sigma)$$

$$prob(signo = 2) = N(-0,5; E[G_t^{a,b}]; \sigma)$$

4.3 Probabilidad del resultado exacto de un partido

A partir de la probabilidad de una determinada diferencia de goles, podemos extraer la probabilidad de un resultado exacto. Para ello, basta con calcular la distribución de frecuencias de los diferentes resultados correspondientes a cada diferencia de goles.

De esta manera, la probabilidad de un determinado resultado $R \{x - y\}$ es igual a la probabilidad de la diferencia de goles d (siendo, por tanto, $d = x - y$) multiplicada por la frecuencia relativa con que se da dicho resultado cuando la diferencia de goles es igual a d , que denotaremos como $frec(x - y | d)$

$$prob(R = x - y) = prob(G_t^{a,b} = d) \times frec(x - y | d).$$

4.4 Probabilidad de una competición

Más allá del resultado de un partido, en ocasiones lo que deseamos conocer es la probabilidad conjunta de todos los posibles resultados de una competición, ya que ello nos permite responder a preguntas como qué probabilidad tiene un determinado equipo de quedar en un puesto concreto, de pasar de fase, de clasificarse para otra competición, de ascender o descender de categoría, etc.

El cálculo exacto de estas probabilidades es inabordable. Incluso si nos ceñimos a un subconjunto reducido de partidos, por ejemplo en un grupo de un mundial formado por cuatro equipos, y para simplificar tomamos sólo los 15 resultados más frecuentes, las combinaciones que se pueden dar con esos resultados en los 6 partidos del grupo son de más de 11 millones, y aún así nos estaríamos dejando resultados por analizar. Si ampliamos la complejidad del problema a una liga de 20 equipos, en la que se juegan 380 partidos, el número de combinaciones supera la capacidad máxima de un ordenador convencional.

Por ello, para calcular este tipo de probabilidades, recurrimos al método de simulación de Monte Carlo, que consiste realizar un número suficientemente alto de simulaciones de los partidos como para hacer converger las probabilidades resultantes a números estables en sus primeros decimales. Generalmente, para alcanzar un grado de convergencia aceptable en una liga de 38 equipos o en una liguilla de un grupo de un mundial formado por cuatro equipos basta con realizar unas 25.000 simulaciones en el primer caso, y unas 10.000 en el segundo.

Una vez realizadas las simulaciones, si la competición o la fase simulada es de tipo liga –en la que todos los equipos se enfrentan entre sí–, la única dificultad está en el almacenar los resultados exactos de todos los partidos y los puntos obtenidos al final por cada equipo en cada simulación, para aplicar los criterios de desempate por los que se rija la competición en los casos en los que dos o más equipos finalicen empatados a puntos.

4.5 Ejemplos numéricos

Para ilustrar las fórmulas anteriores utilizaremos el mundial de Sudáfrica 2010.

En el partido Uruguay vs. Francia correspondiente al grupo A del Mundial de Sudáfrica, las fuerzas de ambos equipos antes de comenzar el encuentro son:

$$F_{t-1}^{URU} = 2,62330$$

$$F_{t-1}^{FRA} = 3,17178$$

El resultado esperado, conforme a la ecuación [8], es:

$$E[G_t^{a,b}] = 2,62330 - 3,17178 + 0 = -0,54848$$

La probabilidad de que la diferencia de goles en dicho encuentro sea, por ejemplo, de +2 goles es aproximadamente de un 8,15%:

$$prob(G_t^{a,b} = 2) = N(2,5, -0,54848, 1,78) - N(1,5, -0,54848, 1,78) \cong 0,08151$$

En este caso, las fuerzas de Uruguay y Francia se actualizarían de la siguiente manera:

$$F_t^{URU} = F_{t-1}^{URU} + (1 - \theta)e_{t-1} = 2,62330 + 0,02 \times (2 - (-0,54848)) = 2,67427$$

$$F_t^{FRA} = F_{t-1}^{FRA} + (1 - \theta)e_{t-1} = 3,17178 - 0,02 \times (2 - (-0,54848)) = 3,12081$$

Como puede observarse, Uruguay ve incrementada notablemente su fuerza, ya que un resultado favorable por 2 goles es muy poco probable, dado que Francia tiene una fuerza superior.

Por otro lado, la probabilidad de que, por ejemplo, gane Francia (equipo visitante), en este caso Francia, es de un 50,6%:

$$prob(signo = 2) = N(-0,5; -0,54848; 1,78) \cong 0,5061$$

En cuanto al resultado exacto, si la diferencia de goles es de +2, y dado que la frecuencia con que se da un resultado de 3-1 condicionada a esa diferencia es de un 29.63%, la probabilidad con la que puede suceder ese resultado es de:

$$prob(R = 3 - 1) = prob(G_t^{a,b} = 2) \times frec(3 - 1 | 2) = 0,08151 \times 0,2963 \cong 0,02415.$$

Por último, la probabilidad de que Francia se clasifique para la siguiente ronda es de un 65,1%, ya que las probabilidades de quedar primero o segundo de grupo son, respectivamente, de un 37,8% y un 27,3%, tras la realización 10.000 simulaciones.