PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1169 2007
TAKSONOMIA 14
Klasyfikacja i analiza danych – teoria i zastosowania

Andrzej Sokołowski
Akademia Ekonomiczna w Krakowie

THE FOOTBALL DISTANCE

„Each match can be won, lost or drawn"
„A match is won by that team which scores more goals"
Kazimierz Górski
(Poland 1974 World Cup Team Manager)

The result of a football match is both qualitative and quantitative. Each match can be won, drawn or lost – and this is the quantitative part, while the number of goals scored and lost is the qualitative one. If we have many guesses or prediction of the forthcoming match and after the match we would like to measure distances between predicted results and the actual one. The classical distance measure (e.g. Minkowski distance) takes into account only the quantitative part. Here is the proposal how we can include also the qualitative part. In fact, in the prediction process it is more important to predict which team will be the winner (or a draw). Number of goals scored and lost is somehow a second choice.

The main assumption of the proposed distance measure is that a result which is correct in qualitative part is closer to the reference one, than any other result different in qualitative part. This can be achieved by shifting compared results to a new three-part, two-dimensional space. The transformation formulas are as follows:

$$\begin{aligned}
&\textit{if} \quad g_H > g_A \rightarrow \begin{cases} g_H' = g_H + 2g_m \\ g_A' = g_A \end{cases} \\
&\textit{if} \quad g_H = g_A \rightarrow \begin{cases} g_H' = g_H + g_m \\ g_A' = g_A + g_m \end{cases} , \\
&\textit{if} \quad g_H < g_A \rightarrow \begin{cases} g_H' = g_H \\ g_A' = g_A + 2g_m \end{cases}
\end{aligned}$$

where:

g_H – numer of goals scored by the home team,

g_A – number of goals scored by the away team,

$g_m = \max\{g_{H_i}, g_{A_i}\}$ - the biggest number of goals scored by a team in a single match found over the set of matches being compared.

Table 1. Distances calculated for $g_m = 1$

	$g_m = 1$			1:0 3:0	0:0 1:1	1:1 2:2	0:1 0:3		
				3	1	2	0	H	
				0	1	2	3	A	
1:0	3:0	3	0	0,00	2,24	2,24	4,24		
0:0	1:1	1	1	2,24	0,00	1,41	2,24		Euclidean
1:1	2:2	2	2	2,24	1,41	0,00	2,24		
0:1	0:3	0	3	4,24	2,24	2,24	0,00		
		Home	Away						
				0	5	5	18		
				5	0	2	5		Sq.eucl.
				5	2	0	5		
				18	5	5	0		
				0	3	3	6		
				3	0	2	3		Manhattan
				3	2	0	3		
				6	3	3	0		

Source: own calculations.

In Table 1 we give an example of distance matrices calculated for the case when the biggest number of goals scored by a single team is 1. We have four possible results which are given in the first row and the first column. Then they are transformed into results given in the second row and the second column. Three distance matrices: Euclidean, squared Euclidean and Manhattan gives of course the same order of distances.

The proposed idea has some wider applicability than just football. In various research, we have many variables, which carry a "double" information, about the quality and about the size. We provide here three examples of such variables. The first one is the actual football game prediction, the second one deals with famous Likert's scale, and the third one – is about below-within-above interval norm of some medical attribute.

Example 1

Seven people tried to predict the result of Manchester United – Celtic Glasgow match played in Champions League group competition in September 2006. The match ended with Manchester 3 – Glasgow 2 result. Whose result was the closest one to the actual score?

Table 2. Calculations for Example 1

Person	Predicted score	Transformed score	Home	Away	Squared euclide-an distance	Rank
A	1:0	13:0	13	0	8	3
B	3:0	15:0	15	0	4	2
C	6:0	18:0	18	0	13	4
D	0:2	0:14	0	14	369	7
E	2:3	2:15	2	15	338	6
F	3:3	9:9	9	9	85	5
G	2:1	14:1	14	1	2	1
Actual score	*3:2*	*15:2*	*15*	*2*	*0*	

gm=6

Source: own calculations.

The closest prediction came from person G.

Example 2

Likert's scale is widely used in psychology and marketing. The subject is given a statement that is a judgment of value rather than a judgment of fact. These statements have to do with wants, desires, conative dispositions of the subjects, not with their opinions regarding matters of fact. After each of these value statements were five responses the subject could choose from: Strongly Disapprove, Disapprove, Neutral/Undecided, Approve, Strongly Approve. How the answers should be coded?

We have three states of attitude towards judged values: positive, neutral and negative. The difference between Strongly Disapprove and Disapprove lies in the "power" of disapproval, but both judgements are the same in "qualitative part". Instead of classical 1 to 5 scoring we should apply the following numerical codes:

1 = Strongly Disapprove
2 = Disapprove
4 = Neutral/Undiceded
6 = Approve
7 = Strongly Approve

The above coding is equivalent to g_M=1, one-dimensional case.

Example 3

The number of White Blood Cells considered as normal varies in 4,500-10,000 range cells/mcl (cells per microliter) [Medline Plus Encyclopedia]. Low numbers of WBCs (leukopenia) may indicate: bone marrow failure (for example, due to infection, tumor, fibrosis), presence of cytotoxic substance, collagen-vascular diseases disease of the liver or spleen, radiation. High numbers of WBCs (leukocytosis) may indicate: infectious diseases, inflammatory disease, leukemia, severe emotional or physical stress, tissue damage (for example, burns), anemia. How should we transformed the number of cells variable to ensure leukopenia

patients are not mixed (in terms of distance) with leukocytosis, and both groups with normal cases?

Normal range equals to: 10000-4500 = 5500.

According to *the football distance idea* normal patients result should be increased by 5500, and leukocytosis patients by 2*5500 = 11000.

It is obvious that recoding of the original values in a football distance manner imposes clustering into three qualitative groups. But this should be treated as an advantage. Two objects from the same qualitative class are always closer to each other than to any other object from other classes. Quality comes before quantity.

ODLEGŁOŚĆ FUTBOLOWA

Streszczenie

W pracy zaproponowano sposób mierzenia odległości między dwoma wynikami meczów piłkarskich. Wynik meczu może być rozpatrywany (z punktu widzenia jednej z drużyn) jako zmienna trzywymiarowa. Pierwszy wymiar ma charakter jakościowy: zwycięstwo, remis, porażka. Dwa następne to zmienne ilościowe: liczba goli strzelonych i liczba goli straconych. W prognozowaniu wyników meczów piłkarskich ważniejsze wydaje się trafne odgadnięcie wyniku „jakościowego", który jednocześnie determinuje relację pomiędzy dwoma wymiarami ilościowymi. Odległość między dwoma wynikami definiowania jest w specjalnej trzysegmentowej przestrzeni dwuwymiarowej. Podano wzory przekształcające surowe wyniki w takie, które pozwalają na zastosowanie klasycznych sposobów liczenia odległości, a jednocześnie zapewniają większe znaczenie jakościowej zgodności wyników.

Przedstawiono też dwa inne zastosowania postulowanego podejścia.