Similarities upper bounds

April 24, 2017

1 Similarities bounds

Recall that we defined the generic average similarities as such :

$$
\begin{aligned}
sim \,:\, & 2^{\mathcal{E}} \times 2^{\mathcal{U}} \times 2^{\mathcal{U}} \longrightarrow [0,1] \\
& (E, G_1, G_2) \longmapsto sim(E, G_1, G_2) = \frac{1}{|E|} \sum_{e \in E} simobj(e, G_1, G_2)
\end{aligned} \tag{1}
$$

Where simobj can be defined depending on the application domain and describe the similarity between two groups of pairs based on their outcome over a given object e.

1.1 (LB^1_{sim},UB^1_{sim}) bounds

We start first by the couple (LB^1_{sim},UB^1_{sim}). Below the definition of the two bounds.

$$LB^1_{sim}(E_c, G_1, G_2) = max\left(\frac{\sigma_{\mathcal{E}} - |E_c|(1 - sim(E_c, G_1, G_2))}{\sigma_{\mathcal{E}}}, 0\right)$$

$$UB^1_{sim}(E_c, G_1, G_2) = min\left(\frac{|E_c| * sim(E_c, G_1, G_2)}{\sigma_{\mathcal{E}}}, 1\right)$$

Recall that $\sigma_{\mathcal{E}}$ define a threshold on the size of subgroup E_c corresponding to a description c over an item.

Given two description c, d where d is a specialization of c : $c \sqsubset d$. We have $E_d \subseteq E_c$. Thus $\sum_{e \in E_d} simobj(e,i,j) \leq \sum_{e \in E_c} simobj(e,i,j) \Leftrightarrow |E_d| sim(E_d,i,j) \leq |E_c| sim(E_c,i,j)$ where i,j are two given groups of individuals.

Proof. UB^1_{sim} :
$sim(E_d,i,j) = \frac{|E_d| sim(E_d,i,j)}{|E_d|} \leq \frac{|E_c| sim(E_c,i,j)}{|\sigma_{\mathcal{E}}|} \Longleftrightarrow$
$|E_d| \geq \sigma_{\mathcal{E}}$ and $|E_d| sim(E_d,i,j) \leq |E_c| sim(E_c,i,j)$, $Q.E.D.$ □

Proof. LB^1_{sim} :
$sim(E_d,i,j) = \frac{|E_d| - |E_d|(1 - sim(E_d,i,j))}{|E_d|} \geq \frac{\sigma_{\mathcal{E}} - |E_c|(1 - sim(E_c,i,j))}{\sigma_{\mathcal{E}}} \Longleftrightarrow$
$\sigma_{\mathcal{E}}[|E_d| - |E_d|(1 - sim(E_d,i,j))] \geq |E_d|[\sigma_{\mathcal{E}} - |E_c|(1 - sim(E_c,i,j))] \Longleftrightarrow$
$\sigma_{\mathcal{E}}|E_d|(1 - sim(E_d,i,j)) \leq |E_d||E_c|(1 - sim(E_c,i,j)) \Longleftrightarrow$
Yet, we have $\sigma_{\mathcal{E}} \leq |E_d|$, thus :
$|E_d|(1 - sim(E_d,i,j)) \leq |E_c|(1 - sim(E_c,i,j)) \Longleftrightarrow$
$|E_d| \sum_{e \in E_d}(1 - simobj(e,i,j)) \leq |E_c| \sum_{e \in E_c}(1 - simobj(e,i,j)) \Longleftrightarrow$
$|E_d| \sum_{e \in E_d}(1 - simobj(e,i,j)) \leq |E_c|[\sum_{e \in E_d}(1 - simobj(e,i,j)) + \sum_{e \in (E_d \setminus E_c)}(1 - simobj(e,i,j))] \Longleftrightarrow$

We denote the quantity $\sum_{e \in E_d}(1 - simobj(e,i,j))$ by α and $\sum_{e \in (E_d \setminus E_c)}(1 - simobj(e,i,j))$ by β. We have $\beta \geq 0$ because $simobj(e,i,j) \in [0,1]$. Thus we write:
$|E_d|\alpha \leq |E_c|[\alpha + \beta] \iff |E_d|\alpha \leq |E_c|\alpha + |E_c|\beta$
Yet $|E_d| \leq |E_c| \Leftrightarrow |E_d|\alpha \leq |E_c|\alpha$ and $|E_c|\beta$ is a positive quantity. $Q.E.D$ $\square$

1.2 (LB^2_{sim},UB^2_{sim}) bounds

Recall below the definition of these two bounds :

$$LB^2_{sim}(E, G_1, G_2) = \frac{1}{\sigma_{\mathcal{E}}} smallest(\{simobj(e, G_1, G_2) \mid e \in E\}, \sigma_{\mathcal{E}})$$

$$UB^2_{sim}(E, G_1, G_2) = \frac{1}{\sigma_{\mathcal{E}}} largest(\{simobj(e, G_1, G_2) \mid e \in E\}, \sigma_{\mathcal{E}})$$

where $smallest(S, n)$ (*resp.* $largest(S, n)$) computes the sum of the n *minimum* (*resp. maximum*) of given set S of real values.

Given two description c, d where d is a specialization of c : $c \sqsubset d$, and i,j two groups of individuals. The proofs of these upper bounds are straight forward. We have $E_d \subseteq E_c$ thus :

$$\frac{1}{\sigma_{\mathcal{E}}} smallest(\{simobj(e,i,j) \mid e \in E_c\}, \sigma_{\mathcal{E}}) \leq \frac{1}{\sigma_{\mathcal{E}}} smallest(\{simobj(e,i,j) \mid e \in E_d\}$$

and

$$\frac{1}{\sigma_{\mathcal{E}}} largest(\{simobj(e,i,j) \mid e \in E_d\}, \sigma_{\mathcal{E}}) \leq \frac{1}{\sigma_{\mathcal{E}}} largest(\{simobj(e,i,j) \mid e \in E_c\}$$

and we have while $|E_d| \geq \sigma_{\mathcal{E}}$ it is obvious that:

$$sim(E_d, i, j) = \frac{1}{|E_d|} \sum_{e \in E_d} simobj(e,i,j) \geq \frac{1}{\sigma_{\mathcal{E}}} smallest(\{simobj(e,i,j) \mid e \in E_d\}, \sigma_{\mathcal{E}})$$

and

$$sim(E_d, i, j) = \frac{1}{|E_d|} \sum_{e \in E_d} simobj(e,i,j) \leq \frac{1}{\sigma_{\mathcal{E}}} largest(\{simobj(e,i,j) \mid e \in E_d\}, \sigma_{\mathcal{E}})$$

Thus the (LB^2_{sim},UB^2_{sim}) are valid.