Quality measures upperbounds

Appendix for ECML/PKDD 2017 paper - *Flash points: Discovering exceptional pairwise behaviors in vote or rating data*

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 $e \in E$.

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 $c \in \mathcal{D}$.

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

Proof. UB^1_{sim} :
$|E_d| \ge \sigma_{\mathcal{E}}$ and $|E_d| sim(E_d,i,j) \le |E_c| sim(E_c,i,j) \Longleftrightarrow$
$sim(E_d,i,j) = \frac{|E_d| sim(E_d,i,j)}{|E_d|} \le \frac{|E_c| sim(E_c,i,j)}{|\sigma_{\mathcal{E}}|}$ and $sim(E_d,i,j) \le 1 \Longleftrightarrow$
$sim(E_d,i,j) \le min\left(\frac{|E_c| sim(E_c,G_1,G_2)}{\sigma_{\mathcal{E}}}, 1\right) = UB^1_{sim}(E_c,G_1,G_2)$, *Q.E.D.* □

Proof. LB^1_{sim} :
$sim(E_d,i,j) = \frac{|E_d| - |E_d|(1 - sim(E_d,i,j))}{|E_d|} \ge \frac{\sigma_{\mathcal{E}} - |E_c|(1 - sim(E_c,i,j))}{\sigma_{\mathcal{E}}} \Longleftrightarrow$
$\sigma_{\mathcal{E}}[|E_d| - |E_d|(1 - sim(E_d,i,j))] \ge |E_d|[\sigma_{\mathcal{E}} - |E_c|(1 - sim(E_c,i,j))] \Longleftrightarrow$
$\sigma_{\mathcal{E}} |E_d|(1 - sim(E_d,i,j)) \le |E_d||E_c|(1 - sim(E_c,i,j)) \Longleftrightarrow$
Yet, we have $\sigma_{\mathcal{E}} \le |E_d|$, thus :
$|E_d|(1 - sim(E_d,i,j)) \le |E_c|(1 - sim(E_c,i,j)) \Longleftrightarrow$
$|E_d| \sum_{e \in E_d}(1 - simobj(e,i,j)) \le |E_c| \sum_{e \in E_c}(1 - simobj(e,i,j)) \Longleftrightarrow$
$|E_d| \sum_{e \in E_d}(1 - simobj(e,i,j)) \le |E_c|[\sum_{e \in E_d}(1 - simobj(e,i,j)) + \sum_{e \in (E_d \setminus E_c)}(1 - simobj(e,i,j))] \Longleftrightarrow$
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 $\beta \ge 0$ because $simobj(e,i,j) \in [0,1]$. Thus we write:
$|E_d|\alpha \le |E_c|[\alpha + \beta] \Longleftrightarrow |E_d|\alpha \le |E_c|\alpha + |E_c|\beta$
Yet $|E_d| \le |E_c| \Leftrightarrow |E_d|\alpha \le |E_c|\alpha$ and $|E_c|\beta$ is a positive quantity. *Q.E.D* □

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 :

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

and

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

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

$$\begin{aligned} sim(E_d, i, j) = \frac{1}{|E_d|} \sum_{e \in E_d} simobj(e, i, j) &\geq \frac{1}{\sigma_{\mathcal{E}}} smallest(\{simobj(e, i, j) \mid e \in E_d\}, \sigma_{\mathcal{E}}) \\ &\geq \frac{1}{\sigma_{\mathcal{E}}} smallest(\{simobj(e, i, j) \mid e \in E_c\}, \sigma_{\mathcal{E}}) = LB^2_{sim}(E_c, i, j) \end{aligned}$$

and

$$\begin{aligned} sim(E_d, i, j) = \frac{1}{|E_d|} \sum_{e \in E_d} simobj(e, i, j) &\leq \frac{1}{\sigma_{\mathcal{E}}} largest(\{simobj(e, i, j) \mid e \in E_d\}, \sigma_{\mathcal{E}}) \\ &\leq \frac{1}{\sigma_{\mathcal{E}}} largest(\{simobj(e, i, j) \mid e \in E_c\}, \sigma_{\mathcal{E}}) = UB^2_{sim}(E_c, i, j) \end{aligned}$$

Thus we have $\forall (c, d) \in \mathcal{D}^2 \mid c \sqsubset d \ : \ LB^2_{sim}(E_c, i, j) \leq sim(E_d, i, j) \leq UB^2_{sim}(E_c, i, j)$ Thus the (LB^2_{sim}, UB^2_{sim}) are valid.

2 Quality measures upper bounds

2.1 Upper bound for $\varphi_{dissent}$:

The quality $\varphi_{dissent}$ measure formula is given by

$$\varphi_{dissent}(d, g', g'') = \frac{\sum_{(i,j) \in \gamma_L(U_{g'}) \times \gamma_L(U_{g''})} max\left(sim\left(E_*, i, j\right) - sim\left(E_d, i, j\right), 0\right)}{|\gamma_L(U_{g'})| . |\gamma_L(U_{g''})|}$$

With $\gamma_L(U_{g'})$ and $\gamma_L(U_{g''})$ two partition of respectively $U_{g'}$ and $U_{g''}$ (g', g'' are two description over the individuals description space, c is a description over the objects description space). We $\gamma_L(U_{g'})$ and $\gamma_L(U_{g''})$ respectively by P_1 and P_2. We rewrite $\varphi_{dissent}(c, g', g'')$ as follows:

$$\varphi_{dissent}(d, g', g'') = \frac{\sum_{(i,j) \in P_1 \times P_2} max\left(sim\left(E_*, i, j\right) - sim\left(E_d, i, j\right), 0\right)}{|P_1| . |P_2|}$$

We have $\forall (c, d) \in \mathcal{D}^2 \,|\, c \sqsubseteq d \ : \ LB_{sim}(E_c, i, j) \leq sim(E_d, i, j) \leq UB_{sim}(E_c, i, j)$ Thus: $max(sim(E_*, i, j) - sim(E_d, i, j), 0) \leq max(sim(E_*, i, j) - LB_{sim}(E_c, i, j), 0)$

Thus : $\varphi_{dissent}(d, g', g'') \leq \frac{\sum_{(i,j) \in P_1 \times P_2} max(sim(E_*, i, j) - LB_{sim}(E_c, i, j), 0)}{|P_1| . |P_2|} = UB_{dissent}(c, g', g'')$

2.2 Upper bound for $\varphi_{consent}$:

The quality $\varphi_{consent}$ measure formula is given by

$$\varphi_{consent}(d,g',g'') = \frac{\sum_{(i,j)\in\gamma_L(U_{g'})\times\gamma_L(U_{g''})} max\,(sim\,(E_d,i,j) - sim\,(E_*,i,j),0)}{|\gamma_L(U_{g'})|.|\gamma_L(U_{g''})|}$$

With $\gamma_L(U_{g'})$ and $\gamma_L(U_{g''})$ two partition of respectively $U_{g'}$ and $U_{g''}$ (g', g'' are two description over the individuals description space, c is a description over the objects description space). We $\gamma_L(U_{g'})$ and $\gamma_L(U_{g''})$ respectively by P_1 and P_2. We rewrite $\varphi_{consent}(c,g',g'')$ as follows:

$$\varphi_{consent}(d,g',g'') = \frac{\sum_{(i,j)\in P_1\times P_2} max\,(sim\,(E_d,i,j) - sim\,(E_*,i,j)\,,0)}{|P_1|.|P_2|}$$

We have $\forall(c,d) \in \mathcal{D}^2\,|\,c \sqsubseteq d\ :\ LB_{sim}(E_c,i,j) \le sim(E_d,i,j) \le UB_{sim}(E_c,i,j)$ Thus: $max(sim(E_d,i,j) - sim(E_*,i,j),0) \le max(UB_{sim}(E_d,i,j) - sim(E_*,i,j),0)$

Thus :

$$\varphi_{dissent}(d,g',g'') \le \frac{\sum_{(i,j)\in P_1\times P_2} max(UB_{sim}(E_c,i,j) - sim(E_*,i,j),0)}{|P_1|.|P_2|} = UB_{consent}(c,g',g'')$$