Standardized Mutual Information for Clustering Comparisons: One Step Further in Adjustment for Chance

Simone Romano SIMONE.ROMANO@UNIMELB.EDU.AU
James Bailey BAILEYJ@UNIMELB.EDU.AU
Nguyen Xuan Vinh VINH.NGUYEN@UNIMELB.EDU.AU
Karin Verspoor KARIN.VERSPOOR@UNIMELB.EDU.AU
Department of Computing and Information Systems, The University of Melbourne, Victoria, Australia

SUPPLEMENTARY MATERIAL

Here we prove Theorem 3 and 4 of Section 3.3 in the paper.

Proof of Theorem 3

Proof. We compute $P(n_{ij})$ only for the lower limit of the hypergeometric random variable support, i.e. $\max\{0, a_i + b_j - N\}$. In both cases $P(n_{ij})$ can be computed in $\mathcal{O}(\max\{a_i, b_j\})$. All other probabilities are computed iteratively and thus their time expense is constant.

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\mathcal{O}(\max\{a_i,b_j\}) + \sum_{n_{ij}=0}^{\min\{a_i,b_j\}} \mathcal{O}(1)\right) = \sum_{i=1}^{r}\sum_{j=1}^{c}\mathcal{O}(\max\{a_i,b_j\}) = \sum_{i=1}^{r}\mathcal{O}(\max\{ca_i,N\})$$
$$= \mathcal{O}(\max\{cN, rN\})$$

□

Proof of Theorem 4

Proof. Each summation over cell values can be bounded above by the maximum value of the cell marginals and each sum can be done in constant time. Let us focus at the inner summations in $E(\text{MI}^2)$:

$$\sum_{j'=1}^{c}\sum_{n_{ij'}=0}^{\max\{a_i,b_{j'}\}}\sum_{i'=1}^{r}\sum_{n_{i'j'}=0}^{\max\{a_{i'},b_{j'}\}}\mathcal{O}(1) = \sum_{j'=1}^{c}\sum_{n_{ij'}=0}^{\max\{a_i,b_{j'}\}}\mathcal{O}(\max\{N, rb_{j'}\})$$
$$= \sum_{j'=1}^{c}\mathcal{O}(\max\{a_iN, a_irb_{j'}, b_{j'}N, rb_{j'}^2\})$$
$$= \mathcal{O}(\max\{ca_iN, a_irN, rN^2\})$$

The above term is thus the computational complexity of the inner loop. Using the same machinery one can prove that:

$$\sum_{j=1}^{c}\sum_{i=1}^{r}\sum_{n_{ij}=0}^{\max\{a_i,b_j\}}\mathcal{O}(\max\{ca_iN, a_irN, rN^2\}) = \mathcal{O}(\max\{c^2N^3, rcN^3\})$$

□