Supplementary document to Theoretical and empirical analyses of fixed-size redescription set construction

Matej Mihelčić[1 [0000−0002−1023−8413]]

Faculty of Science, University of Zagreb, Zagreb, Croatia matmih@math.hr

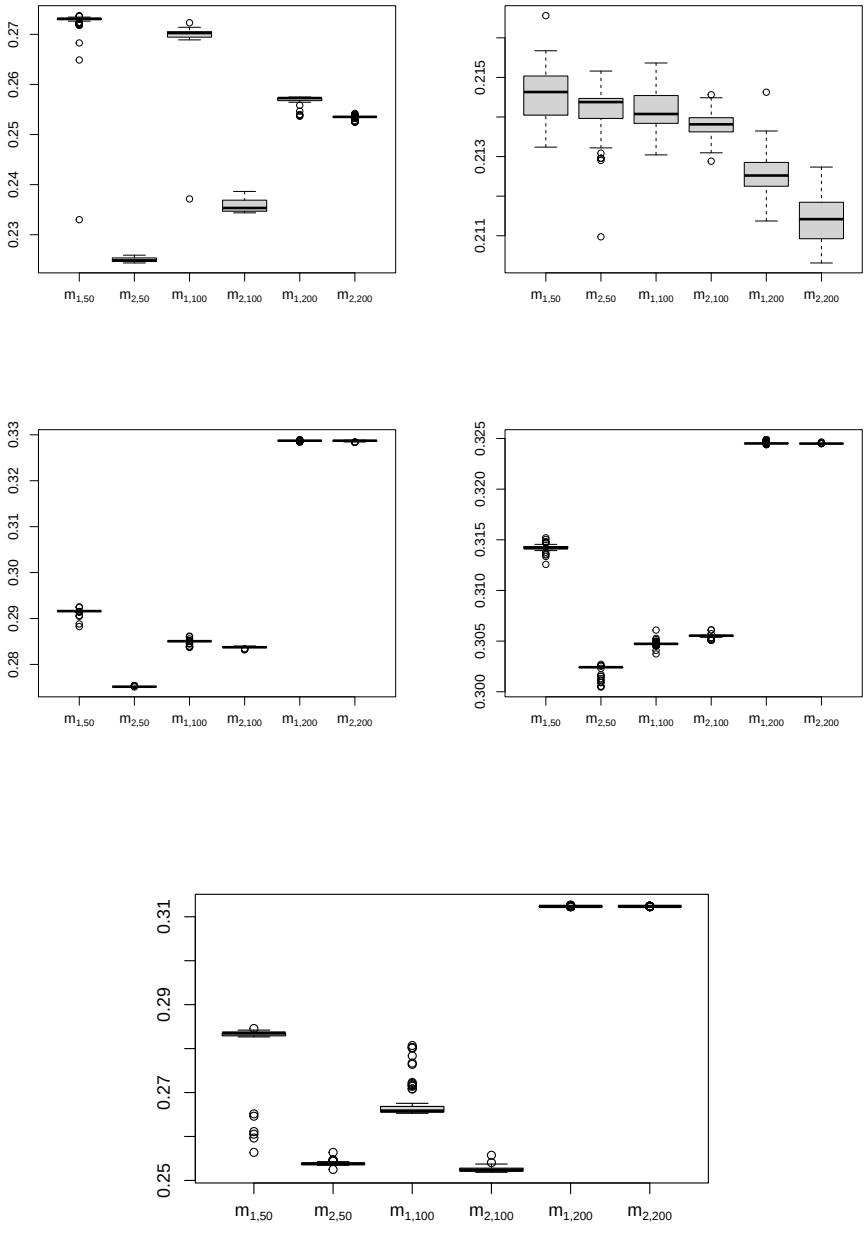
1 Variability of the output redescription set scores dependent on the seed redescription

Distributions of the output redescription scores, obtained utilizing updates mes_1 and $mesU_2$, as described in the main manuscript, can be seen in Figure 1. Results obtained on the Country, the Phenotype and the Bio datasets directly follow the intuition that the redescription sets obtained using updates $mesU_1$ have larger variability than sets obtained using $mesU_2$. Also, variability decreases with the increase of the output redescription set size. Results on the DBLP dataset follow the same notion, except the output sets of size 100, where sets using mes_1 mostly achieve better score than these using $mesU_2$. This can be the result of the specifics of the dataset, where specifically optimizing the instance/attribute coverage can significantly increase redundancy. On the Mammals dataset, we do not observe the reduction in variability for the sets of size 200, however $|\mathcal{R}|_{all,200} = 3054$ is substantially larger than $|\mathcal{R}_{all,50}| = 2413$ and $\mathcal{R}_{all,100} = 1969$ which might have influenced the end distribution of output redescription set scores for $m = 200$.

2 CLUS-RM parameters used to obtain $\mathcal{R}_{all}$

In all experiments, we used Predictive Clustering regression trees of depth 8. In addition to the main search guiding trees, we used a supplementary random forest with 6 trees. We utilized redescription query minimization, the redescription refinement procedure and allowed the use of logical conjunction, disjunction and negation operator to construct redescription queries. Maximum redescription support set size was set to $\approx 0.8 \cdot |E|$ on all datasets. Minimal redescription support size was set to 5 on the Country, Phenotype (dataset with small $|E| < 200$) and the DBLP dataset (sparse data) and on 10 on the Mammals dataset (dataset with medium-sized $|E| < 5000$) and the Bio dataset ($|E| < 4000$). Maximal redescription p-value of 0.01 was used in all experiments. The minimal redescription Jaccard index of 0.7 was used on the Bio dataset, 0.5 was used on the Country, the Phenotype and the Mammals dataset and 0.3 on the sparse DBLP dataset. We used 5 random restarts (initializations) on the Country and the DBLP datasets, 6 on the Phenotype dataset and 2 on the Mammals and the

Fig. 1. Distributions of the *mes* score of the m candidate sets obtained by utilizing each redescription from $\mathcal{R}_{mes_1}$ and $\mathcal{R}_{mes_2}$ as a seed redescription on the Country (top-left), the Mammals (top-right), the Phenotype (middle-left), the DBLP (middle-right) and the Bio (bottom) datasets. $m_{1,50}$ denotes measure distribution of a set obtained using mes_1 of a size 50, and $m_{2,50}$ denotes measure distribution of the set obtained using $mesU_2$ of the same size.



Bio dataset. 20 iterations per a restart was used on the Country, 50 on the Phenotype, 5 on the DBLP and the Bio and 2 on the Mamamls dataset.

3 Execution times

Local search procedure execution times can be seen in Table 1. It should be noted that the algorithm creates 2 sequences of candidate redescription set, where each sequence contains $(m+1)$ candidate sets that are iteratively improved ($m \in \{50,\ 100\}$). 3 runs are used for candidates of size 50 and 4 for candidates of size 100, and for each run, the number of iterations was $4^k \cdot 10$. Thus, for each sequence of sets of size 50, we create 42891 candidate sets, and for each sequence of sets of size 100, we create 343501 candidate sets. The total number is 85782 candidates for sets of size 50 and 687002 candidates for sets of size 100.

Table 1. TW_i denotes the attribute type of the i-th view (N - numeric, B - Boolean). ET_s denotes execution times on redescription sets of size s.

$\mathcal{D}$	$\lvert E\rvert$	$\lvert V\rvert$	T_{W_1}	T_{W_2}	$\lvert\mathcal{R}_{all}\rvert$	ET_{50}	ET_{100}
Country	199	361	N	N	{3262, 3176}	9h, 48m, 47s	3d, 20h, 7m, 35s
Mammals	2575	242	B	N	{2413, 1969}	10h, 35m, 10s	4d, 2h, 55m, 22s
Phenotype	92	1230	N	N	{530, 585}	4h, 30m, 5s	2d, 0h, 36m, 42s
DBLP	6455	6759	B	B	{292, 188}	11h, 10m, 5s	2d, 3h, 1m, 45s
Bio	3475	4523	N	N	{3412, 2075}	4d, 2h, 17m, 46s	3w, 5d, 12h, 1m, 10s

Experiments were performed using 15 computation threads on the `Intel(R) Xeon(R) W-1370 @ 2.90GHz`.