Order Is All You Need for Categorical Data Clustering

Author Response to Reviewers Qzf2 and RcUz
Results on Larger-Scale Categorical Datasets

Anonymous Author(s)

Six additional larger datasets in different fields have been chosen from the UCI machine learning repository (see http://archive.ics.uci.edu/ml) to further evaluate the clustering performance and execution time of the proposed OCL method. The statistics of the newly added datasets are summarized in Table 1. All datasets are pre-processed following the same settings in the submitted paper (see Section 4.1 and Appendix A.1).

Clustering Performance. The clustering performance of different methods is compared on the six additional datasets in Table 2. Consistent with the results in our paper, the proposed OCL outperforms the other methods in general. More detailed observations are provided below: (1) Average performance ranking of OCL is around 1.4 in 18 comparisons formed on 6 datasets under 3 metrics, while the most competitive rivals, i.e., DLC and H2H, rank around 5.7 and 4.4, respectively. The huge ranking gap (larger than 3) intuitively demonstrates the superiority of OCL. (2) All the experimental datasets are real public datasets from various fields, such as healthcare, gaming, recommendation systems, finance, etc. This configuration is sufficient to verify the superiority of OCL without bias. (3) The proposed OCL always ranks first or second in all comparisons (except for NMI performance on the BM dataset) and performs the best in 13 out of 18 comparisons, indicating that the advantage of OCL is very stable and extensive.

Execution Time. To evaluate the efficiency of OCL, Table 3 reports the average execution time of each compared method on different datasets corresponding to the above clustering performance experiment. It can be observed that on the three large-scale datasets, i.e., CT4, BM, and CR, the execution time of OCL is shorter than or in the same magnitude as the advanced UDMC, DLC, and H2H methods. As for the other three medium-sized datasets, we record the running time on them for completeness, and it is not very meaningful to discuss efficiency on them. It is noteworthy that the most effective efficiency verification is in our submitted paper. We have analyzed that the time complexity of OCL is linear to n, and implemented experiments in Section 4.4 to evaluate its execution time variation trend on datasets of different sizes (ranging from 10k to 100k). They all consistently prove that OCL is highly scalable to large-scale datasets and does not incur much extra computational cost compared to the existing advanced counterparts.

Table 1: Dataset statistics. $d^{\{\text{nom}\}}$, $d^{\{\text{ord}\}}$, n, and k^* are the numbers of nominal attributes, ordinal attributes, samples, and true clusters, respectively. The full name of the "Coupon" dataset is "In-Vehicle Coupon Recommendation".

No.	Dataset	Abbrev.	$d^{\{\text{nom}\}}$	$d^{\{\text{ord}\}}$	n	k^*
1	Obesity Levels	OB	6	2	2111	7
2	Auction Verifications	AV	0	6	2043	2
3	Chess	CC	36	0	3195	2
4	Coupon	CR	7	4	12684	2
5	Bank Marking	BM	6	3	45211	2
6	Connect-4	CT4	42	0	67557	3

Table 2: Clustering performance evaluated by CA, NMI, and ARI. The best and second-best results on each dataset are highlighted in bold and underlined, respectively. The "$\overline{AR}$" row reports the average performance ranks of each method in all 18 comparisons.

Data	Metric	KMD	LSM	JDM	CBDM	UDMC	DLC	H2H	HDC	ADC	OCL
CT4	CA	0.4000±0.02	0.3609±0.01	0.3696±0.02	0.3651±0.02	0.3819±0.03	0.3618±0.01	0.3924±0.03	0.3879±0.03	0.3914±0.03	**0.4121±0.02**
	ARI	-0.0027±0.00	0.0010±0.00	0.0001±0.00	0.0004±0.00	0.0013±0.00	0.0012±0.00	0.0006±0.00	-0.0000±0.00	-0.0005±0.00	**0.0150±0.01**
	NMI	0.0028±0.00	0.0029±0.00	0.0017±0.00	0.0015±0.00	0.0036±0.00	0.0013±0.00	0.0011±0.00	0.0016±0.00	0.0016±0.00	**0.0128±0.01**
OB	CA	0.3715±0.03	0.3697±0.02	0.3594±0.03	0.3673±0.03	0.3720±0.02	0.3724±0.04	0.3797±0.02	0.3617±0.04	0.3765±0.03	**0.3935±0.01**
	ARI	0.1527±0.03	0.1588±0.02	0.1347±0.04	0.1522±0.04	0.1580±0.02	0.1548±0.04	**0.1727±0.02**	0.1417±0.05	0.1472±0.04	0.1696±0.01
	NMI	0.2256±0.03	0.2330±0.02	0.2081±0.04	0.2262±0.03	0.2342±0.02	0.2391±0.03	**0.2644±0.02**	0.2261±0.04	0.2233±0.03	0.2633±0.01
AV	CA	0.6251±0.08	0.6035±0.10	0.6291±0.12	**0.6695±0.10**	0.6065±0.10	0.6150±0.06	0.6441±0.14	0.6364±0.10	0.6325±0.12	0.6637±0.07
	ARI	0.0073±0.02	0.0111±0.02	0.0136±0.03	0.0365±0.02	0.0123±0.03	0.0262±0.03	0.0214±0.03	0.0199±0.02	0.0173±0.02	**0.0393±0.02**
	NMI	0.0021±0.00	0.0024±0.00	0.0024±0.00	0.0099±0.01	0.0027±0.00	0.0093±0.01	0.0075±0.01	0.0035±0.00	0.0034±0.00	**0.0110±0.01**
BM	CA	0.5730±0.06	0.5792±0.04	0.5959±0.07	0.6011±0.07	0.5709±0.03	0.5944±0.05	0.6064±0.12	0.5457±0.03	0.5543±0.04	**0.6127±0.00**
	ARI	0.0146±0.03	0.0107±0.02	0.0183±0.04	0.0054±0.05	0.0074±0.02	0.0178±0.02	0.0186±0.01	-0.0037±0.02	0.0035±0.02	**0.0233±0.00**
	NMI	**0.0211±0.00**	0.0143±0.01	0.0135±0.01	0.0198±0.01	0.0179±0.01	0.0121±0.01	0.0155±0.01	0.0192±0.01	0.0204±0.01	0.0194±0.00
CC	CA	0.5562±0.04	0.5572±0.03	0.5632±0.03	0.5433±0.03	0.5587±0.02	0.5342±0.04	0.5248±0.02	0.5438±0.03	0.5395±0.04	**0.5726±0.05**
	ARI	0.0165±0.01	0.0166±0.01	0.0189±0.02	0.0103±0.01	0.0150±0.01	0.0099±0.02	0.0019±0.01	0.0108±0.01	0.0113±0.02	**0.0285±0.02**
	NMI	0.0150±0.01	0.0149±0.01	0.0149±0.01	0.0084±0.01	0.0119±0.01	0.0076±0.01	0.0023±0.01	0.0148±0.01	0.0141±0.02	**0.0213±0.02**
CR	CA	0.5209±0.02	0.5247±0.02	0.5185±0.02	0.5297±0.02	0.5245±0.01	0.5488±0.03	0.5541±0.00	0.5466±0.02	0.5504±0.02	**0.5590±0.02**
	ARI	0.0021±0.00	0.0033±0.00	0.0024±0.00	0.0039±0.00	0.0030±0.00	0.0120±0.01	0.0114±0.00	0.0101±0.01	0.0106±0.01	**0.0145±0.01**
	NMI	0.0017±0.00	0.0024±0.00	0.0020±0.00	0.0022±0.00	0.0022±0.00	0.0071±0.01	**0.0115±0.00**	0.0065±0.00	0.0057±0.00	0.0083±0.01
$\overline{AR}$		6.44	6.06	6.72	5.78	5.89	5.67	4.44	6.39	6.00	1.39

Table 3: Execution time on the six larger-scale datasets (in seconds).

Data	KMD	LSM	JDM	CBDM	UDMC	DLC	H2H	HDC	ADC	OCL
CT4	1.5167	58.4583	2.0139	13.0987	1588.7899	251.3594	118.8687	4.8770	4.1545	537.1386
BM	0.9939	21.6699	1.5092	5.1621	277.1126	51.8629	36.6258	1.5586	1.5205	221.3714
CR	0.6927	8.9719	1.6845	0.6929	1.7732	10.3699	2.7615	0.4437	0.3136	10.2580
OB	0.0666	0.8645	0.0775	0.0699	0.0839	1.2523	0.8365	0.0278	0.0306	7.9199
AV	0.0272	0.3103	0.0161	0.0109	0.0208	0.0709	0.0633	0.0138	0.0109	0.3194
CC	0.0557	0.9704	0.0946	0.1001	0.0905	4.8657	2.3374	0.0453	0.0689	2.7972