Learning Order Forest for Qualitative-Attribute Data Clustering

Mingjie Zhao[1], Sen Feng[1], Yiqun Zhang[1,6,*], Mengke Li[2,3,6], Yang Lu[4,5,6] and Yiu-Ming Cheung[6]

[1]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China
[2]Guangdong Laboratory of Artifcial Intelligence and Digital Economy (SZ), Shenzhen, China
[3]School of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[4]Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, China
[5]Key Laboratory of Multimedia Trusted Perception and Effcient Computing, Ministry of Education of China, Xiamen University, China
[6]Department of Computer Science, Hong Kong Baptist University, Hong Kong, China

Appendix

A Details of Compared Methods

The traditional K-MoDes algorithm(KMD) [3], tailored for categorical data, served as the baseline. Notable methods include the entropy-based LSM [6] and the Context-Based Distance Metric (CBDM) [4], both of which are representative similarity measures designed for categorical data. These methods were combined with k-modes to create two distinct clustering approaches for comparison. The state-of-the-art clustering methods included Jia's Distance Metric (JDM) [5], iterative clustering learning based on sample-cluster similarity metric OCIL [1], clustering with the Unified Distance Metric (UDMC) [10], HD-based clustering (HDC) [9], Projection-based clustering from Heterogeneous to Homogeneous spaces (H2H) [12], the Distance Learning-based Clustering (DLC) algorithm [8], and clustering with the Graph-Based Dissimilarity Measurement Applicable to Anytype-attributed data clustering (ADC) [11].

B Complementary Experimental Results

B.1 Comparisons of Different Distance Structures

We have conducted an intuitive comparison of different distance structure generation ways, i.e., RGGs, FCGs, RGLGs, and SLGs, on four typical datasets, as shown in Figure S.1 in the submitted paper. For completeness, comparisons on the remainder eight datasets are provided here. Since some of the datasets have natural distance structures closer to a line graph, the performance of RGGs will degenerate to be similar to that of RGLGs, which is still consistent with our hypothesis, i.e., a more flexible distance structure that can represent multiple local order relationships among possible values would be more competent in categorical data clustering.

B.2 Clustering Performance

Clustering performance of different methods are compared in Tables S.1 w.r.t. NMI. The best and second-best results on each dataset are highlighted in **bold** and underline, respectively. The observations include the following three aspects: (1) Overall, COForest performs the best on almost all datasets, indicating its superiority in clustering. (2) The performance of COForest on the ZO, HF, and DS datasets, although the best, is not significantly better than the second-best rivals. However, the second-best method varies on these datasets, indicating the robustness of COForest. (3) Although COForest does not have the best NMI performance on VT and TS datasets, it is not surpassed by much by the winners and still being very competitive. In addition, results of CBDM on CE, NS, and LS datasets are not reported because the attributes of these datasets are independent of each other, making CBDM fail in measuring distances based on the context of attributes.

Significance test results w.r.t. NMI metric is demonstrated in Figure S.2. The test is conducted by first implementing Friedman test on the average performance ranks reported in the last rows in Tables S.1. The corresponding p-values is 0.0025, passing the Friedman test under 95% confidence interval (i.e., $\alpha = 0.05$). On this basis, Bonferroni Dunn (BD) post-hoc test is implemented accordingly. Critical Difference (CD) intervals for the two-tailed BD tests at 95% (α= 0.05) and 90% (α = 0.1) confidence intervals are 3.8048 and 3.5204, respectively, for comparing 11 methods across 12 datasets. Almost all the compared methods fall outside the right boundary of the CD intervals centered at COForest, indicating that COForest significantly outperforms them.

B.3 Clustering Performance on Mixed-Data

To demonstrate the potential of COForest in being extended to mixed data clustering with both numerical and categorical attributes, we combine the distance metric learned on the categorical attributes and then combine it with Euclidean distance metric for k-prototypes clustering. Figure S.3 demonstrates its clustering accuracy against the original k-prototypes [2] and a mix-data clustering method ek-prototypes [7]. It can be observed that COForest can always outperform k-prototypes on the datasets with mixed attributes, demonstrating that the COForest is promising in being applied to more complex mixed data clustering scenarios.

* Corresponding Author. Email: yqzhang@gdut.edu.cn

Table S.1. Clustering performance evaluated by NMI. "$\overline{AR}$" row reports the average performance rankings.

Data	KMD	LSM	JDM	CBDM	OCIL	UDMC	DLC	H2H	HDC	ADC	COForest (ours)
HR	0.0113±0.01	0.0087±0.00	0.0091±0.00	0.0285±0.04	0.0061±0.01	0.0116±0.00	0.0070±0.01	0.0000±0.00	0.0093±0.01	0.0236±0.04	**0.0576±0.05**
CE	0.0525±0.03	0.0712±0.04	0.0721±0.04	-	0.1221±0.08	0.0703±0.04	0.1246±0.04	0.0555±0.05	0.0525±0.03	0.0525±0.03	**0.1827±0.07**
AC	0.2105±0.06	0.2552±0.04	0.1485±0.13	0.1950±0.09	0.2931±0.13	0.2399±0.08	0.2668±0.19	0.2668±0.00	0.2122±0.08	0.2558±0.08	**0.3740±0.11**
VT	0.4477±0.02	0.4667±0.01	0.4667±0.01	0.4923±0.00	**0.4987±0.01**	0.4570±0.01	0.4546±0.16	0.4893±0.00	0.4893±0.00	0.4844±0.00	0.4897±0.00
CS	0.0077±0.01	0.0116±0.03	0.0129±0.03	0.0205±0.01	0.0172±0.03	0.0290±0.03	0.0451±0.03	0.0338±0.01	0.0292±0.02	0.0301±0.02	**0.0814±0.02**
SB	0.8574±0.12	0.8888±0.15	0.7993±0.15	0.8592±0.13	0.8422±0.31	0.9048±0.10	0.9119±0.12	0.9539±0.10	0.8588±0.10	0.8869±0.10	**0.9734±0.08**
NS	0.0742±0.03	0.0705±0.03	0.0631±0.03	-	0.1738±0.14	0.0781±0.04	0.1366±0.11	0.1403±0.12	0.0742±0.03	0.0742±0.03	**0.2110±0.19**
ZO	0.7380±0.05	0.7524±0.06	0.7994±0.06	0.7708±0.04	0.6336±0.34	0.7600±0.07	0.7878±0.05	0.8220±0.04	0.7661±0.07	0.7844±0.06	**0.8233±0.09**
TS	0.0098±0.01	0.0087±0.00	0.0089±0.00	0.0082±0.00	**0.0101±0.00**	0.0082±0.01	0.0074±0.00	0.0095±0.00	0.0068±0.01	0.0065±0.00	0.0072±0.00
HF	**0.0020±0.00**	**0.0020±0.00**	0.0014±0.00	0.0009±0.00	0.0011±0.00	0.0015±0.00	0.0011±0.00	0.0013±0.00	0.0015±0.00	0.0015±0.00	**0.0020±0.00**
DS	0.2014±0.19	0.2014±0.19	0.2134±0.20	0.2513±0.21	0.2997±0.32	0.1710±0.19	0.2997±0.32	0.0697±0.03	0.1710±0.19	0.2134±0.20	**0.3065±0.17**
LS	0.2151±0.11	0.2667±0.13	0.2667±0.13	-	0.2601±0.14	0.2968±0.17	0.2711±0.14	0.2516±0.09	0.2151±0.11	0.2151±0.11	**0.4154±0.16**
$\overline{AR}$	7.6250	6.5417	6.9583	7.3750	5.5000	6.0000	5.0417	5.4167	7.3333	6.2917	1.9167

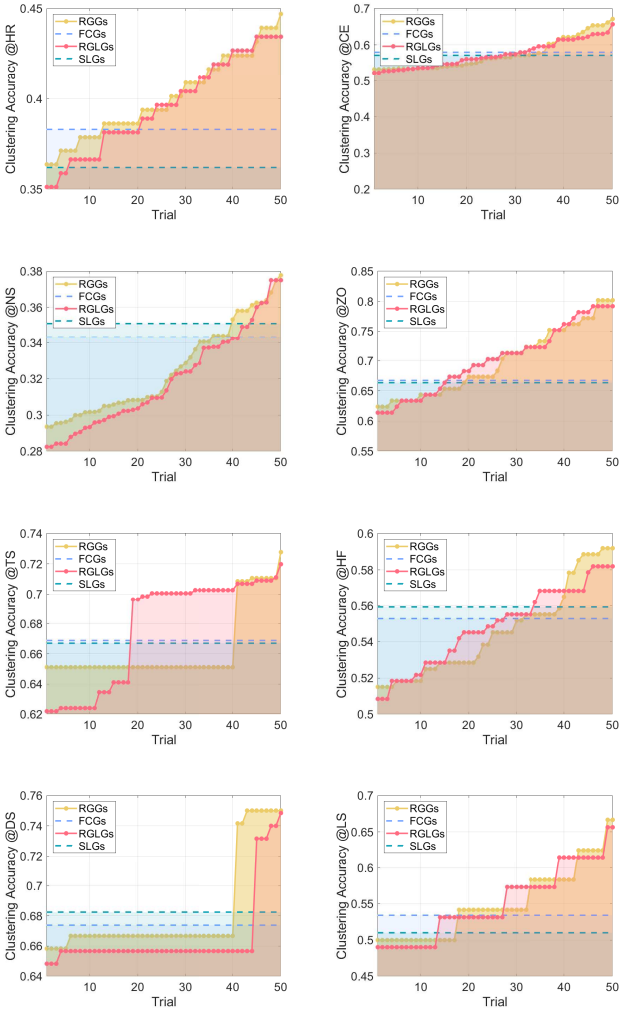

Figure S.1. Supplementary material about toy-example in Figure S.1: performance comparison of other datasets

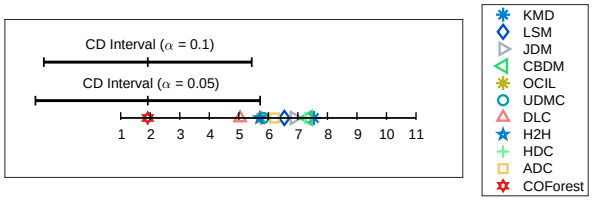

Figure S.2. Results of the two-tailed BD tests w.r.t. the NMI performance shown in Table S.1.

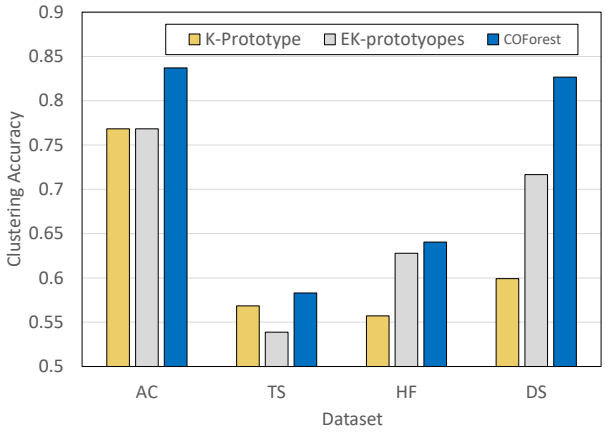

Figure S.3. Comparison of Clustering accuracy performance between k-prototypes, ek-prototypes, and COForest methods on mixed datasets composed of numerical and categorical attributes, i.e., AC, TS, HF, DS in Table 2 in the submitted paper by adding their numerical attributes back. Note that in the submitted paper, we aim to evaluate the clustering performance on categorical data, so the numerical attributes are omitted.

References

[1] Y.-m. Cheung and H. Jia. Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. *Pattern Recognition*, 46(8):2228–2238, 2013.

[2] Z. Huang. Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 21–34, 1997.

[3] Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2 (3):283–304, 1998.

[4] D. Ienco, R. G. Pensa, and R. Meo. From context to distance: Learning dissimilarity for categorical data clustering. *ACM Transactions on Knowledge Discovery from Data*, 6(1):1–25, 2012.

[5] H. Jia, Y.-m. Cheung, and J. Liu. A new distance metric for unsupervised learning of categorical data. *IEEE Transactions on Neural Networks and Learning Systems*, 27(5):1065–1079, 2016.

[6] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, 1998.

[7] R. S. Sangam and H. Om. An equi-biased k-prototypes algorithm for clustering mixed-type data. *SÄądhanÄą*, 43, 2018.

[8] Y. Zhang and Y.-m. Cheung. An ordinal data clustering algorithm with automated distance learning. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 6869–6876, 2020.

[9] Y. Zhang and Y.-m. Cheung. Learnable weighting of intra-attribute distances for categorical data clustering with nominal and ordinal attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3560–3576, 2022.

[10] Y. Zhang and Y.-m. Cheung. A new distance metric exploiting heteroge-

neous inter-attribute relationship for ordinal-and-nominal-attribute data clustering. *IEEE Transactions on Cybernetics*, 52(2):758–771, 2022.

[11] Y. Zhang and Y.-M. Cheung. Graph-based dissimilarity measurement for cluster analysis of any-type-attributed data. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):6530–6544, 2023.

[12] Y. Zhang, Y. ming Cheung, and A. Zeng. Het2hom: Representation of heterogeneous attributes into homogeneous concept spaces for categorical-and-numerical-attribute data clustering. In *International Joint Conference on Artificial Intelligence*, 2022.