CADM: CLUSTER-CUSTOMIZED ADAPTIVE DISTANCE METRIC FOR CATEGORICAL DATA CLUSTERING

Taixi Chen[1] *Yiu-ming Cheung*[2†] *Yiqun Zhang*[2]

[1] Binghamton University, Binghamton, NY, USA
[2] Hong Kong Baptist University, Hong Kong SAR, China

ABSTRACT

An appropriate distance metric is crucial for categorical data clustering, as the distance between categorical data cannot be directly calculated. However, the distances between attribute values usually vary in different clusters induced by their different distributions, which has not been taken into account, thus leading to unreasonable distance measurement. Therefore, we propose a cluster-customized distance metric for categorical data clustering, which can competitively update distances based on different distributions of attributes in each cluster. In addition, we extend the proposed distance metric to the mixed data that contains both numerical and categorical attributes. Experiments demonstrate the efficacy of the proposed method, i.e., achieving an average ranking of around first in fourteen datasets. The source code is available at https://anonymous.4open.science/r/CADM-47D8/

Index Terms— Categorical data, Clustering, Distance metric, Unsupervised learning

1. INTRODUCTION

Cluster analysis of categorical data composed of nominal and ordinal attributes are common in many fields, such as medical analysis, customer questionnaires, and so on [1, 2, 3, 4]. Nevertheless, due to the difficulty of measuring the difference between categorical attributes, the core problem of categorical data clustering relies on discovering and defining a proper distance metric for effective measurement. The existing distance metrics have been explored along two main branches: 1) directly calculate the distance of categorical data based on the defined encoding methods [5, 6, 7, 8], and 2) indirectly estimate the distance between different attributes based on the frequency or distribution in context [9, 10]. However, most of them neglect the heterogeneity between ordinal and nominal attributes in categorical data.

Recently, the order information of ordinal data has received increasing attention [11, 12, 13, 14] because order reflects the intrinsic difference between ordinal attribute values. For instance, in the Nursery categorical dataset (Table 1 (I)), its ordinal attribute $social$ contains three ordinal values: $nonprob$, $slightly\_prob$, and $problematic$. The distance between $nonprob$ and $problematic$ should not only consider their difference, as their semantic concepts are not isolated and independent, but related to their medium value $slightly\_prob$ as well [15].

However, existing methods consider the order information, the intrinsic distance between ordinal attribute values, to be the same in the entire dataset, ignoring the heterogeneity of different clusters. It is not reasonable in many cases, for example, in Table 1 (II), the distance between $nonprob$ and $problematic$ in the class spec-prior and priority is larger than that in class not-recom, based on their frequency in context, because their importance is different in different classes. Moreover, Table 1 (II) also shows that the distance of nominal attribute values varies in three classes induced by the distinction of their frequency distributions in different classes. Thus, it can be observed from the two sub-figures that the commonly used total context frequency distribution cannot reflect this distribution difference for both nominal and ordinal attribute values between different clusters, which restricts the performance of the distance metric and clustering.

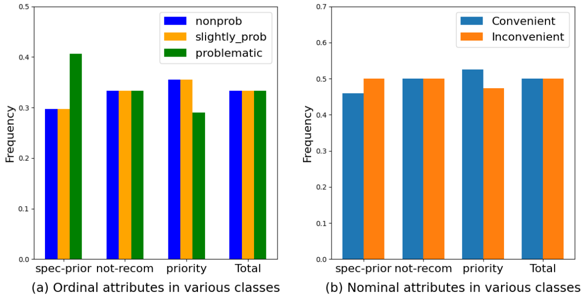
To tackle these challenges, this paper proposes a novel distance metric called Cluster-customized Adaptive Distance Metric (CADM). This is a unified distance metric for both ordinal and nominal data. It defines the attribute value distances between objects and different cluster centers as Cluster-customized attribute Value Distance (CVD), depending on Cluster-customized Value Importance (CVI), adaptively changing in different clusters during iterations. Specifically, CVI is the importance of one attribute value in different clusters, which is determined by both this attribute value's count and the maximum count of all attribute values in this attribute. Based on CVD, the data with high cluster importance attribute values will be pulled closer to the cluster center, as it represents this cluster. Otherwise, it will be pulled away from the cluster center. The intuition behind this idea is to reasonably leverage information guidance from different clusters to improve distance measurement. Based on the observation of Table 1, it is necessary to design more refined distance measurements for attribute values.

Furthermore, a Cluster-customized Attribute Importance (CAI) is defined to weigh attribute contributions in forming

[†] Corresponding author: ymc@comp.hkbu.edu.hk

ID	finance	social	health
People_1	convenient	nonprob	recommended
People_2	convenient	nonprob	priority
People_3	convenient	nonprob	not_recom
People_4	convenient	slightly_prob	recommended
People_5	convenient	slightly_prob	priority
People_6	convenient	problematic	recommended
People_7	convenient	problematic	priority
People_8	convenient	problematic	not_recom
People_9	inconv	nonprob	recommended
People_10	inconv	nonprob	priority

(I)



(a) Ordinal attributes in various classes

(b) Nominal attributes in various classes

(II)

Table 1. (I) The examples in the Nursery categorical dataset. The finance attribute is nominal, while others are ordinal. (II) The distance and distribution of both ordinal and nominal attributes are different in each cluster.

distances, which regards the consistency of possible attribute values in one attribute category. This mechanism is applicable in attribute-independent cases as it depends on the self-importance. In addition, we extend CADM to mixed data with heterogeneous attributes in the experiment. In summary, this paper makes the following contributions:

- A unified distance metric CADM is proposed for nominal and ordinal data that considers the adaptive cluster-customized distance measurement, addressing the problem of distance difference in various clusters.
- Based on the CVI, the CVD is defined to dynamically measure the attribute value distance between categorical data and the cluster center. It can provide personalized measurements for each cluster, reducing bias during the clustering process.
- To weigh the attribute contributions in forming distances, this paper defines the CAI, which can achieve minute adjustments to CVD, making distance measurement more reasonable and accurate.

2. PROPOSED METHOD

In this section, we first formulate the problem. Then, we introduce our proposed distance metrics CADM and provide the algorithm analysis of it.

2.1. Problem Formulation

Assuming a dataset S can be rewritten as $S =< X, A, O >$. The data object sets $X = [x_0, x_1, ..., x_{n-1}]$ with n objects. And for each sample $x_i = [x_i^0, x_i^1, ..., x_i^{d-1}]$ because it has the d attributes. Moreover, as the attribute $A = [A^0, A^1, ..., A^{d-1}]$, for each attribute, it has n values so that $A^r = [A_0^r, A_1^r, ..., A_{n-1}^r]$. Besides, as each attribute A^r must have limited possible values v^r, thus, the unique set O^r for different attribute A^r can be written as $O^r = [o_0^r, o_1^r, ..., o_{v^r-1}^r]$, which is ascending order. Besides, data should be placed as $A = A^{num} + A^{nom} + A^{ord}$, while the numerical data A^{num} is optional. It is worth noting that mixed and categorical data clustering normally adopts the k-prototypes clustering algorithm [16, 12, 17], which only considers the distance between categorical data and cluster centers. Each cluster is described by a center $c_l = [c_l^0, c_l^1, ..., c_l^{d-1}]$ from $C = [c_0, c_1, ..., c_{k-1}]$. It aims to assign n data objects in X to k proper clusters, which can be formulated as minimizing:

$$J = \sum_{i=0}^{n-1} \sum_{l=0}^{k-1} d(x_i, c_l) \tag{1}$$

where c_l is one specific cluster center, and the value of c_l^r is the most frequent possible value from A^r in l_{th} cluster. The dissimilarity between an object and the cluster center can be rewritten as

$$d(x_i, c_l) = \sum_{r=0}^{d-1} d_m(x_i^r, c_l^r) + d_I(A^r), \tag{2}$$

where the $d_m(.)$ is the distance between the categorical attribute values, and $d_I(.)$ measures the importance of the categorical attribute A^r.

2.2. Cluster-customized Adaptive Distance Metric

CADM is proposed to adaptively measure the cluster-personalized distance between categorical data. Thus, we define the distances of different categorical attribute values, such as x_i^r and c_l^r, as computed by:

$$d_m(x_i^r, c_l^r) = \begin{cases} \sum_{j=\min(\alpha(x_i^r, c_l^r))}^{\max(\alpha(x_i^r, c_l^r))} d_a^l(o_j^r, o_p^r), & A^r \in A^{ord} \\ d_a^l(o_t^r, o_p^r), & A^r \in A^{nom} \end{cases} \tag{3}$$

where x_i^r and c_l^r are denoted as o_t^r and o_p^r in O^r perspective. The o_j^r represents the intermediate attribute value between x_i^r and c_l^r, including x_i^r as well. If attribute A^r is an ordinal attribute, CADM uses order information from the intermediate

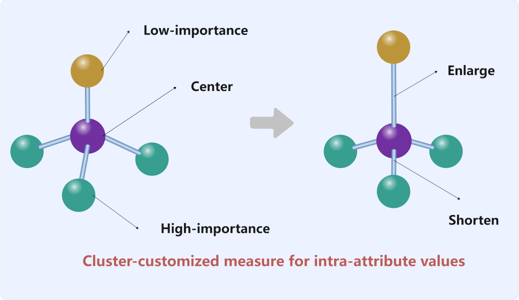


Fig. 1. Framework of attribute value distance measurement

attribute value following existing works [12, 14] to enhance measurement. The $d_a^l(.)$ is the CVD designed for measuring the distance of attribute values. The notation l means distance measure in the l_{th} cluster. The $\alpha(x_i^r, c_l^r)$ fetches the order number of x_i^r and c_l^r from O^r. In addition, the $d_m(x_i^r, c_l^r)$ is defined as zero when $x_i^r = c_l^r$, and so is $d_a^l(.)$.

Furthermore, as shown in Fig 1, the CVD is proposed to cluster-customized measure the distance between attribute values, which is defined as:

$$d_a^l(o_s^r, o_p^r) = \gamma^l(o_s^r) + \gamma^l(o_p^r), \tag{4}$$

where the o_s^r is used to represent o_j^r and o_t^r for generality reasons, and it is called the rival attribute value. $\gamma^l(.)$ is the rival factor for attribute values. It is designed to construct the CVD based on the CVI of both cluster center and categorical data for reasonable measurement.

Remark 1 *As Fig 1 shows, $d_m(.)$ is a cluster-customized measure that can adaptively change in different clusters. CVD defines that the data with high CVI is closer to the cluster center, as it represents this cluster. Otherwise, it should be far away from the cluster center. This can be regarded as a rival process between different o_s^r and o_p^r in each cluster. Thus, we define a rival factor as a bridge between the CVD and CVI to achieve this rival process for measuring the distance of attribute values.*

Specifically, we define the rival factor as:

$$\gamma^l(o_z^r) = \begin{cases} CVI^l(o_p^r), & o_z^r \in o_p^r \\ \frac{1}{CVI^l(o_s^r)}. & o_z^r \in o_s^r \end{cases} \tag{5}$$

Since the cluster center's attribute value (o_p^r) is generally of high importance, its rival factor should contribute greatly to the distance based on CVI, and it is the basis of the rival process. Moreover, the rival factor for the rival attribute value (o_s^r) is a reciprocal form of the cluster center's rival factor $\gamma^l(o_p^r)$. The intuition is that the rival factor of the rival attribute value will enlarge the distance when the rival attribute value's CVI is low. The CVI is computed by:

$$CVI^l(o_s^r) = \frac{C^l(o_s^r)}{\max\limits_{1 \le f \le v^r} C(o_f^r)}, \tag{6}$$

Algorithm 1 Proposed CADM clustering algorithm

Input: Dataset **S**, number **k** of clusters
Output: cluster label **L**

```
while $C^t \neq C^{t-1}$ do
    Calculate the distance between attribute values based on the CVD by Eq.(4).
    Utilizing Eq.(8) to obtain the attribute importance to further constrain the distance.
    Gaining final distance measurement based on Eq.(2).
    Update D, C and L.
end while
return L
```

which offers the relative importance of the intra-attribute value. The $C^l(.)$ provides the count of one attribute value in a specific cluster. Differently, $C(.)$ provides the total counts of one attribute value in the whole dataset S. Thus, CVI is the frequency count of the rival attribute value. In this case, the CVI will adaptively update in each iteration and cluster. Moreover, the importance of different attribute categories varies, and they are different in other clusters. Thus, CAI is defined to explicitly weigh the contributions of attribute categories in forming distances, which is computed by:

$$CAI^l(A^r) = \frac{\max\limits_{1 \le s \le v^r} C^l(o_s^r)}{n}, \tag{7}$$

Thus, CAI can be leveraged to define attribute importance, which is computed by:

$$d_I(A^r) = CAI^l(A^r)^2. \tag{8}$$

Remark 2 *CAI is calculated as a cohesion factor considering the consistency of the count of the possible attribute values in specific attribute A^r. Thus, the higher the maximum count of possible attribute values, the more consistent the attribute, and the more weight will be added to the distance calculated in this attribute.*

The whole algorithm process is shown in **Algorithm 1**. It adopts the framework of K-modes clustering that iteratively updates cluster center C, distance matrices D, and cluster labels L in each time step t until convergence.

3. EXPERIMENT

Nine Counterparts: including three classical (i.e., HDM [5], GSM [18], LSM [19]), two context-based (CBDM [20], EBDM [21]), and four SOTA (i.e.,UDM [12], HARR [17], COF [22], QGRL [23]) clustering algorithm are chosen. Especially, QGRL is a deep learning based algorithm, while others are unsupervised learning. Set the cluster number k as the real class number of the data. We ran ten times for each experiment and used the average value in the report table.

Table 2. Experiments with competitive distance metric in categorical, ordinal, and nominal datasets. ”−” indicates that the algorithm is inapplicable or has not converged in one dataset.

Dataset Statistics			HDM [5]	GSM [18]	LSM [19]	CBDM [20]	EBDM [21]	UDM [12]	HARR [17]	COF [22]	QGRL [23]	**CADM**
Abbrev.	#Instance	k	Baseline	Baseline	Baseline	2012	2020	2022	2025	2024	2024	Ours
NS	12960	4	0.375 ± 0.04	0.356 ± 0.03	0.375 ± 0.03	−	0.400 ± 0.02	$\underline{0.411 \pm 0.03}$	0.407 ± 0.02	0.362 ± 0.09	0.395 ± 0.02	$\mathbf{0.429 \pm 0.03}$
PR	123	12	0.410 ± 0.04	0.393 ± 0.04	0.396 ± 0.04	0.399 ± 0.03	0.361 ± 0.04	0.412 ± 0.03	0.431 ± 0.06	0.429 ± 0.03	$\mathbf{0.678 \pm 0.02}$	$\underline{0.433 \pm 0.05}$
HA	132	3	0.389 ± 0.02	0.398 ± 0.04	0.392 ± 0.04	0.383 ± 0.04	0.407 ± 0.03	0.446 ± 0.04	0.447 ± 0.03	$\underline{0.453 \pm 0.02}$	0.362 ± 0.02	$\mathbf{0.471 \pm 0.03}$
LY	148	4	0.459 ± 0.05	0.451 ± 0.04	0.459 ± 0.05	0.489 ± 0.05	0.450 ± 0.03	$\underline{0.494 \pm 0.03}$	0.453 ± 0.04	0.488 ± 0.12	0.462 ± 0.03	$\mathbf{0.507 \pm 0.04}$
SM	61069	2	0.506 ± 0.01	0.508 ± 0.01	$\underline{0.530 \pm 0.01}$	−	0.520 ± 0.02	0.521 ± 0.01	0.516 ± 0.02	0.504 ± 0.02	−	$\mathbf{0.550 \pm 0.03}$
C4	67577	3	0.371 ± 0.03	0.373 ± 0.03	0.358 ± 0.01	−	0.356 ± 0.04	0.378 ± 0.02	0.383 ± 0.03	$\mathbf{0.431 \pm 0.03}$	−	$\underline{0.411 \pm 0.03}$
VT	435	2	0.874 ± 0.01	0.534 ± 0.01	0.534 ± 0.00	0.806 ± 0.01	0.853 ± 0.00	0.872 ± 0.00	0.873 ± 0.01	0.875 ± 0.01	$\mathbf{0.884 \pm 0.02}$	$\underline{0.880 \pm 0.00}$
LS	24	3	0.375 ± 0.02	0.502 ± 0.01	$\underline{0.595 \pm 0.03}$	0.515 ± 0.01	0.508 ± 0.02	0.550 ± 0.03	0.501 ± 0.03	0.563 ± 0.08	−	$\mathbf{0.608 \pm 0.03}$
PE	101	7	0.485 ± 0.03	0.515 ± 0.03	0.419 ± 0.02	0.409 ± 0.02	$\underline{0.610 \pm 0.03}$	0.609 ± 0.02	0.545 ± 0.03	0.561 ± 0.04	0.557 ± 0.03	$\mathbf{0.615 \pm 0.04}$
LE	1000	5	0.269 ± 0.04	0.298 ± 0.03	0.303 ± 0.03	0.306 ± 0.02	0.369 ± 0.02	$\underline{0.372 \pm 0.03}$	0.345 ± 0.04	0.319 ± 0.06	0.337 ± 0.02	$\mathbf{0.373 \pm 0.02}$
AA	104	2	0.577 ± 0.01	0.510 ± 0.02	0.576 ± 0.02	−	0.601 ± 0.03	0.567 ± 0.03	0.560 ± 0.02	0.559 ± 0.04	$\underline{0.636 \pm 0.01}$	$\mathbf{0.661 \pm 0.03}$
HF	299	2	0.599 ± 0.02	0.679 ± 0.01	0.602 ± 0.02	−	0.625 ± 0.03	0.600 ± 0.02	0.704 ± 0.03	0.692 ± 0.02	$\underline{0.713 \pm 0.03}$	$\mathbf{0.736 \pm 0.03}$
HD	297	5	0.351 ± 0.02	0.358 ± 0.04	0.391 ± 0.04	−	0.360 ± 0.03	0.377 ± 0.04	0.417 ± 0.03	0.403 ± 0.04	$\underline{0.432 \pm 0.02}$	$\mathbf{0.471 \pm 0.03}$
MM	824	2	0.818 ± 0.00	0.820 ± 0.00	0.831 ± 0.00	0.828 ± 0.00	0.807 ± 0.00	$\mathbf{0.837 \pm 0.00}$	0.818 ± 0.00	0.826 ± 0.01	0.830 ± 0.00	$\underline{0.832 \pm 0.00}$
		Rank:	7.1	7.7	6.6	7.0	6.1	3.9	4.9	4.6	$\underline{3.0}$	**1.3**

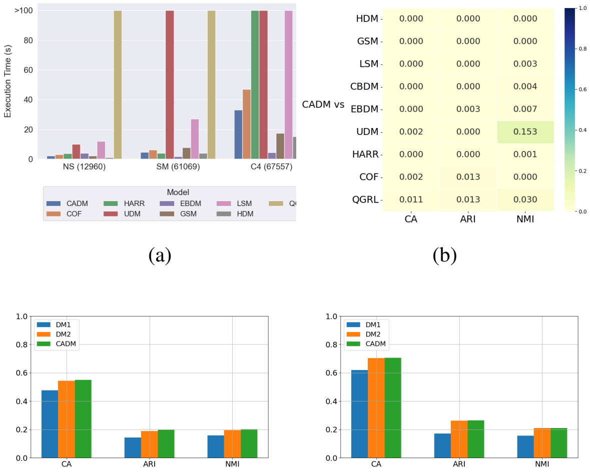
Fourteen Datasets: are collected from [24] and [25] shown in Table 2, including 4 mixed (i.e., AA, HF, HD, MM), 5 categorical (i.e., NS, PR, HA, LY, SM), 3 ordinal (i.e., C4, LE, LS), and 2 nominal datasets (i.e., VT, PE).

Validity Indices: Clustering Accuracy (CA) [26] is selected for evaluating the clustering performance. Larger values indicate better clustering performance.

Comparative Results: The bold value means the best performance in a dataset, and the underline value means the second-best performance in one dataset. As Table 2 shows, the proposed CADM outperforms nine counterparts with an average rank 1.3, indicating its superiority in categorical and mixed data clustering. On categorical datasets (i.e., NS, LY, SM), the advantage of CADM is extremely obvious, indicating that the proposed cluster-personalized metric can provide more accurate distance information for each cluster. The superiority of CADM on the mixed dataset (i.e., AA, HF, HD) is also tremendous, which illustrates its significant universality in heterogeneous datasets. Moreover, Fig 2 (b) shows the results of the Wilcoxon signed rank test between our CADM and the other nine methods in fourteen datasets, which indicates CADM has a significant superiority over other methods, achieving a 95% confidence level.

Efficiency Evaluation : We select three large datasets (i.e., NS, SM, C4) to examine model efficiency. Based on the Fig 2 (a), CADM outperforms the four latest SOTA models. Although three baselines are faster, their clustering performance is extremely lower than CADM in fourteen datasets.

Ablation studies: In ablation studies, DM1 is a simple distance measurement leveraging order information. DM2 adds the CVD, and CADM adds the CAI. The results in Fig 2 (c) and (d) illustrate the effectiveness of the proposed cluster-customized meteic framework. Specifically, it is obvious that CVD drastically improves the performance, indicating the benefits of the cluster-customized framework, and CAI also effectively adjusts the final measurement. More comparative results (in other indicators), complexity analysis, and proofs can be found in online appendix.



(a) (b)

(c) (d)

Fig. 2. (a) efficiency test on three large datasets. (b) Wilcoxon signed rank test in fourteen datasets. (c) and (d) demonstrate the ablation study in the categorical and mixed datasets.

4. CONCLUDING REMARKS

This paper proposes a novel cluster-customized adaptive distance metric for categorical data clustering. It is a unified distance metric for categorical data, which is applicable to both nominal and ordinal data. Specifically, Cluster-customized attribute value distance measurement is defined considering the competitive cluster-customized strategy to address the concern of the distance difference between two attribute values in various clusters. Besides, the importance of the attribute has been proposed to weigh the contributions of different attributes in forming distance, making the distance measurement more reasonable. Experiments have shown CADM's superiority in categorical data clustering. Moreover, it is efficient without any pre-set parameters, and its mechanisms have high interpretability, indicating its significant potential.

5. REFERENCES

[1] Alan Agresti, *Categorical data analysis*, vol. 792, John Wiley & Sons, 2012.

[2] Tiago RL Dos Santos and Luis E Zárate, "Categorical data clustering: What similarity measure to recommend?," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1247–1260, 2015.

[3] Giulia Caruso, SA Gattone, Francesca Fortuna, and Tonio Di Battista, "Cluster analysis for mixed data: An application to credit risk evaluation," *Socio-Economic Planning Sciences*, vol. 73, pp. 100850, 2021.

[4] Duy-Tai Dinh, Van-Nam Huynh, and Songsak Sriboonchitta, "Clustering mixed numerical and categorical data with missing values," *Information Sciences*, vol. 571, pp. 418–442, 2021.

[5] Floriana Esposito, Donato Malerba, V Tamma, HH Bock, et al., "Classical resemblance measures," *Studies In Classification, Data Analysis, and Knowledge Organization*, vol. 15, pp. 139–152, 2000.

[6] Si Quang Le and Tu Bao Ho, "An association-based dissimilarity measure for categorical data," *Pattern Recognition Letters*, vol. 26, no. 16, pp. 2549–2557, 2005.

[7] Yuhua Qian, Feijiang Li, Jiye Liang, Bing Liu, and Chuangyin Dang, "Space structure and clustering of categorical data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 10, pp. 2047–2059, 2015.

[8] Songlei Jian, Guansong Pang, Longbing Cao, Kai Lu, and Hang Gao, "Cure: Flexible categorical data representation by hierarchical coupling learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 5, pp. 853–866, 2018.

[9] Hong Jia, Yiuming Cheung, and Jiming Liu, "A new distance metric for unsupervised learning of categorical data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 5, pp. 1065–1079, 2015.

[10] Markus Ring, Florian Otto, Martin Becker, Thomas Niebler, Dieter Landes, and Andreas Hotho, "Condist: A context-driven categorical distance measure," in *Proceedings of Machine Learning and Knowledge Discovery in Databases*. Springer, 2015, pp. 251–266.

[11] Yiqun Zhang and Yiuming Cheung, "Exploiting order information embedded in ordered categories for ordinal data clustering," in *Proceedings of Foundations of Intelligent Systems: 24th International Symposium*. Springer, 2018, pp. 247–257.

[12] Yiqun Zhang and Yiuming Cheung, "A new distance metric exploiting heterogeneous interattribute relationship for ordinal-and-nominal-attribute data clustering," *IEEE Transactions on Cybernetics*, vol. 52, no. 2, pp. 758–771, 2022.

[13] Yiqun Zhang and Yiuming Cheung, "An ordinal data clustering algorithm with automated distance learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 6869–6876.

[14] Yiqun Zhang and Yiuming Cheung, "Learnable weighting of intra-attribute distances for categorical data clustering with nominal and ordinal attributes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3560–3576, 2021.

[15] Pengkai Wang, Yunfan Zhang, Yiqun Zhang, Yang Lu, Mengke Li, and Yiuming Cheung, "Clustering by learning the ordinal relationships of qualitative attribute values," in *Proceedings of 2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[16] Zhexue Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[17] Yiqun Zhang, Mingjie Zhao, Yizhou Chen, Yang Lu, and Yiuming Cheung, "Learning unified distance metric for heterogeneous attribute data clustering," *Expert Systems with Applications*, p. 126738, 2025.

[18] David W Goodall, "A new similarity index based on probability," *Biometrics*, pp. 882–907, 1966.

[19] Dekang Lin et al., "An information-theoretic definition of similarity.," in *Proceedings of International Conference on Machine Learning*, 1998, vol. 98, pp. 296–304.

[20] Amir Ahmad and Lipika Dey, "A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set," *Pattern Recognition Letters*, vol. 28, no. 1, pp. 110–118, 2007.

[21] Yiqun Zhang, Yiuming Cheung, and Kay Chen Tan, "A unified entropy-based distance metric for ordinal-and-nominal-attribute data clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 1, pp. 39–52, 2020.

[22] Mingjie Zhao, Sen Feng, Yiqun Zhang, Mengke Li, Yang Lu, and Yiuming Cheung, "Learning order forest for qualitative-attribute data clustering," in *ECAI 2024*, pp. 1943–1950. IOS Press, 2024.

[23] Junyang Chen, Yuzhu Ji, Rong Zou, Yiqun Zhang, and Yiuming Cheung, "Qgrl: quaternion graph representation learning for heterogeneous feature data clustering," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 297–306.

[24] Ian H Witten, Eibe Frank, Mark A Hall, Christopher J Pal, and Mining Data, "Practical machine learning tools and techniques," in *Proceedings of Data mining*. Elsevier Amsterdam, The Netherlands, 2005, vol. 2, pp. 403–413.

[25] Pawel Bielski and Dustin Kottonau, "Micro gas turbine electrical energy prediction," UCI Machine Learning Repository, 2024, https://doi.org/10.24432/C58S4T.

[26] Xiaofei He, Deng Cai, and Partha Niyogi, "Laplacian score for feature selection," *Advances in Neural Information Processing Systems*, vol. 18, 2005.