

CLUSTER-CUSTOMIZED ADAPTIVE DISTANCE METRIC FOR CATEGORICAL DATA CLUSTERING

A. APPENDIX

A.1. Datasets and Counterparts

Table 1. Statistics of fifteen datasets, where A(O), A(N), and A(U) indicate the number of ordinal attributes, nominal attributes, and numerical attributes, respectively. Moreover, attribute type "Mixed" represents datasets containing both categorical and numerical attributes.

Attribute type	Dataset	Instance	A(O)	A(N)	A(U)	Cluster
Mixed	US	2458285	3	57	9	—
	AA	104	3	4	2	2
	HF	299	0	5	7	2
	HD	297	4	3	6	5
	MM	824	2	2	1	2
Categorical	SM	61069	6	2	0	2
	NS	12960	7	1	0	4
	PR	123	1	15	0	12
	HA	132	2	2	0	3
	LY	148	3	15	0	4
Ordinal	C4	67577	42	0	0	3
	LE	1000	4	0	0	5
	LS	24	4	0	0	3
Nominal	VT	435	0	16	0	2
	PE	101	0	16	0	7

As Table 1 shows, five mixed datasets include US Census (abbreviated as US), Autism-Adolescent (abbreviated as AA), Heart Failure (abbreviated as HF), Heart Disease (abbreviated as HD), and Mammographic Mass (abbreviated as MM), which are benchmark datasets from the UCI Machine Learning Repository (UCIMLR) (Bielski & Kottonau, 2024). Moreover, we select five categorical benchmarks, including Secondary Mushroom (abbreviated as SM), Nursery (abbreviated as NS), Primary Tumor (abbreviated as PR), Hayes Roth (abbreviated as HA), and Lymphography (abbreviated as LY) from UCIMLR (Bielski & Kottonau, 2024). Besides, three ordinal benchmark data containing Connect4 (abbreviated as C4), Lecturer Evaluation (abbreviated as LE), and Lenses (abbreviated as LS) from UCIMLR (Bielski & Kottonau, 2024). The LE is collected from Weka datasets (Witten et al., 2005); C4 and LS are collected from UCIMLR (Bielski & Kottonau, 2024). Two nominal datasets, Photo Evaluation (abbreviated as PE) and Voting Records (abbreviated as VT), are also benchmark datasets from real-world ordinal question-

naires and UCIMLR (Bielski & Kottonau, 2024).

Table 2. Information of the 9 counterparts. "Type" indicates how the methods define their distance metric. QGRL is a deep learning method that is different from the other 8 methods.

No.	Counterpart	Year	Type
1	GSM	1966	Direct
2	LSM	1998	Context-based
3	HDM	2000	Direct
4	CBDM	2012	Context-based
5	EBDM	2020	Weight-based
6	UDM	2022	Weight-based
7	COF	2024	Learning-based
8	QGRL*	2024	Deep learning
9	HARR	2025	Learning-based

In particular, as Table 2 shows, CADM is compared with nine competitive similarity measures. The HDM (Esposito et al., 2000) and GSM (Goodall, 1966) are the traditional direct calculation algorithms. LSM (Lin et al., 1998) and CBDM (Ahmad & Dey, 2007) are the context-based algorithms. EBDM (Zhang et al., 2020) and UDM (Zhang & Cheung, 2022) are similarity measures of weight-based type algorithms, and UDM (Zhang & Cheung, 2022) is regarded as state-of-the-art (SOTA). COF (Zhao et al., 2024), QGRL (Chen et al., 2024), and HARR (Zhang et al., 2025) are regarded as state-of-the-art (SOTA) similarity measures of representative learning-based type algorithms.

A.2. Clustering Performance

In the first group experiment, based on Table 3, we have some observations about our CADM in categorical, ordinal, and nominal data clustering: 1) The proposed CADM outperforms all its counterparts in three metrics across ten datasets, achieving a total ranking of 1.5, which indicates its overall effectiveness in clustering. Moreover, it is worth noting that the total rank of the second-place algorithm, UDM (3.7), is significantly lower than that of CADM (1.5). 2) On most challenging categorical datasets in the first experiment (i.e., NS, PR, LY), CADM outperforms almost all algorithms. This illustrates the superiority of CADM, which is an unsupervised categorical data distance metric, indicating the effectiveness of CVD and CAI in constructing distances for categorical data. 3) On large and high-dimensional datasets (i.e., SM and

Table 3. Experiments with competitive distance metric in categorical, ordinal, and nominal datasets. "—" indicates that the algorithm is inapplicable or has not converged in one dataset.

Index	Dataset	HDM	GSM	LSM	CBDM	EBDM	UDM	HARR	COF	QGRL	CADM
CA	NS	0.375 ± 0.04	0.356 ± 0.03	0.375 ± 0.03	—	0.400 ± 0.02	<u>0.411 ± 0.03</u>	0.407 ± 0.02	0.362 ± 0.09	0.395 ± 0.02	0.429 ± 0.03
	PR	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	0.678 ± 0.02	0.433 ± 0.05
	HA	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	<u>0.453 ± 0.02</u>	0.362 ± 0.02	0.471 ± 0.03
	LY	0.459 ± 0.05	0.451 ± 0.04	0.459 ± 0.05	0.489 ± 0.05	0.450 ± 0.03	<u>0.494 ± 0.03</u>	0.453 ± 0.04	0.488 ± 0.12	0.462 ± 0.03	0.507 ± 0.04
	SM	0.506 ± 0.01	0.508 ± 0.01	<u>0.530 ± 0.01</u>	—	0.520 ± 0.02	0.521 ± 0.01	0.516 ± 0.02	0.504 ± 0.02	—	0.550 ± 0.03
	C4	0.371 ± 0.03	0.373 ± 0.03	<u>0.358 ± 0.01</u>	—	0.356 ± 0.04	0.378 ± 0.02	0.383 ± 0.03	0.431 ± 0.03	—	<u>0.411 ± 0.03</u>
	VT	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	0.884 ± 0.02	<u>0.880 ± 0.00</u>
	LS	0.375 ± 0.02	0.502 ± 0.01	<u>0.595 ± 0.03</u>	0.515 ± 0.01	0.508 ± 0.02	0.550 ± 0.03	0.501 ± 0.03	0.563 ± 0.08	—	0.608 ± 0.03
	PE	0.485 ± 0.03	0.515 ± 0.03	0.419 ± 0.02	0.409 ± 0.02	0.610 ± 0.03	0.609 ± 0.02	0.545 ± 0.03	0.561 ± 0.04	0.557 ± 0.03	0.615 ± 0.04
	LE	0.269 ± 0.04	0.298 ± 0.03	0.303 ± 0.03	0.306 ± 0.02	0.369 ± 0.02	<u>0.372 ± 0.03</u>	0.345 ± 0.04	0.319 ± 0.06	0.337 ± 0.02	0.373 ± 0.02
Rank:		7.1	7.7	6.6	7.0	5.9	3.5	4.8	4.3	<u>3.2</u>	1.3
ARI	NS	0.048 ± 0.02	0.045 ± 0.01	0.058 ± 0.02	—	0.062 ± 0.02	0.064 ± 0.02	0.042 ± 0.05	0.135 ± 0.18	0.016 ± 0.01	<u>0.085 ± 0.04</u>
	PR	0.193 ± 0.03	0.173 ± 0.04	0.184 ± 0.03	0.185 ± 0.03	0.140 ± 0.05	0.196 ± 0.04	0.159 ± 0.05	0.178 ± 0.07	0.491 ± 0.02	<u>0.204 ± 0.01</u>
	HA	−0.030 ± 0.01	0.000 ± 0.01	−0.001 ± 0.02	0.000 ± 0.02	0.008 ± 0.02	0.032 ± 0.02	<u>0.051 ± 0.01</u>	0.043 ± 0.03	0.001 ± 0.02	0.062 ± 0.03
	LY	0.105 ± 0.03	0.106 ± 0.04	0.114 ± 0.04	0.135 ± 0.04	0.210 ± 0.04	0.085 ± 0.03	0.142 ± 0.03	0.112 ± 0.13	0.092 ± 0.02	<u>0.156 ± 0.04</u>
	SM	−0.004 ± 0.02	−0.001 ± 0.01	0.000 ± 0.01	—	0.001 ± 0.03	<u>0.002 ± 0.02</u>	0.001 ± 0.02	0.000 ± 0.02	—	0.010 ± 0.01
	C4	−0.002 ± 0.04	0.000 ± 0.01	0.001 ± 0.01	—	0.001 ± 0.04	0.002 ± 0.02	0.003 ± 0.03	<u>0.006 ± 0.02</u>	—	0.007 ± 0.03
	VT	0.466 ± 0.03	0.469 ± 0.02	0.532 ± 0.03	0.372 ± 0.05	0.498 ± 0.04	0.528 ± 0.03	0.551 ± 0.03	0.561 ± 0.02	<u>0.593 ± 0.03</u>	0.615 ± 0.05
	LS	0.106 ± 0.05	0.001 ± 0.02	0.029 ± 0.01	0.002 ± 0.03	0.050 ± 0.01	0.131 ± 0.02	0.135 ± 0.03	0.336 ± 0.02	—	<u>0.314 ± 0.01</u>
	PE	0.061 ± 0.04	0.001 ± 0.07	−0.002 ± 0.01	−0.001 ± 0.01	<u>0.225 ± 0.03</u>	0.154 ± 0.01	0.007 ± 0.02	0.127 ± 0.06	0.165 ± 0.02	0.226 ± 0.05
	LE	0.043 ± 0.04	0.037 ± 0.03	0.000 ± 0.01	0.038 ± 0.04	<u>0.067 ± 0.03</u>	0.074 ± 0.03	<u>0.094 ± 0.03</u>	0.045 ± 0.05	0.077 ± 0.01	0.099 ± 0.04
Rank:		7.5	8.1	6.9	7.6	4.9	4.4	4.5	<u>4.1</u>	4.7	1.4
NMI	NS	0.051 ± 0.02	0.047 ± 0.02	0.065 ± 0.02	—	0.064 ± 0.02	0.067 ± 0.03	<u>0.083 ± 0.03</u>	0.081 ± 0.12	0.015 ± 0.02	0.084 ± 0.03
	PR	0.345 ± 0.03	0.103 ± 0.02	0.435 ± 0.03	0.438 ± 0.04	0.413 ± 0.03	<u>0.447 ± 0.03</u>	0.318 ± 0.04	0.346 ± 0.06	0.503 ± 0.02	0.347 ± 0.01
	HA	0.016 ± 0.01	0.016 ± 0.02	0.014 ± 0.02	0.014 ± 0.02	0.024 ± 0.02	0.061 ± 0.04	0.066 ± 0.03	<u>0.077 ± 0.02</u>	0.005 ± 0.01	0.081 ± 0.06
	LY	0.171 ± 0.04	0.178 ± 0.06	0.172 ± 0.04	0.185 ± 0.04	0.210 ± 0.04	0.218 ± 0.03	0.161 ± 0.04	0.174 ± 0.09	0.110 ± 0.02	<u>0.216 ± 0.01</u>
	SM	<u>0.006 ± 0.02</u>	0.000 ± 0.01	0.000 ± 0.01	—	0.002 ± 0.02	0.001 ± 0.01	0.001 ± 0.02	0.001 ± 0.02	—	0.007 ± 0.01
	C4	0.001 ± 0.03	0.000 ± 0.02	0.001 ± 0.01	—	0.003 ± 0.04	0.002 ± 0.02	0.001 ± 0.03	<u>0.004 ± 0.03</u>	—	0.006 ± 0.02
	VT	0.510 ± 0.03	0.221 ± 0.04	0.341 ± 0.05	0.372 ± 0.03	0.419 ± 0.02	0.356 ± 0.03	0.557 ± 0.02	<u>0.564 ± 0.02</u>	0.545 ± 0.01	0.583 ± 0.03
	LS	0.134 ± 0.02	0.014 ± 0.01	0.029 ± 0.00	0.021 ± 0.03	0.162 ± 0.03	<u>0.269 ± 0.04</u>	0.136 ± 0.04	0.127 ± 0.07	—	0.359 ± 0.05
	PE	0.063 ± 0.02	0.005 ± 0.01	0.007 ± 0.02	0.002 ± 0.02	0.227 ± 0.03	0.308 ± 0.02	0.231 ± 0.03	0.135 ± 0.03	0.204 ± 0.03	<u>0.261 ± 0.03</u>
	LE	0.005 ± 0.01	0.058 ± 0.03	0.004 ± 0.01	0.038 ± 0.02	0.092 ± 0.01	<u>0.114 ± 0.01</u>	0.074 ± 0.02	0.096 ± 0.02	0.130 ± 0.01	0.108 ± 0.03
Rank:		6.4	8.2	7.3	6.9	4.3	<u>3.3</u>	5.0	4.3	5.7	1.9
Total Rank:		7.0	8.0	6.9	7.2	5.0	<u>3.7</u>	4.7	4.2	4.5	1.5

C4), CADM exhibits the best performance in ARI and NMI metrics, indicating its superiority in constructing a reliable distance for complex attributes. 4) Although QGRL (Chen et al., 2024) achieves better performance on some datasets, it is obvious that the performance of CADM is superior and more stable in general, and does not need extensive weight parameters for learning over a long time in each dataset.

In the second group of experiments in Table 4, CADM still demonstrates its superiority in mixed data, achieving a total ranking of 1.3, which illustrates the effectiveness of CADM in constructing a reasonable distance for categorical data that is homogeneous with the distance of numerical data. Moreover, it is noteworthy that CADM outperforms almost all methods, whatever the number of numerical and categorical attributes, indicating its significant universality in heterogeneous datasets. Although the deep learning method exhibits better performance in the ARI metric on the MM dataset, CADM is more competitive in the CA and NMI metrics, further indicating its efficacy compared with deep learning methods.

A.3. Effectiveness of mCVD

We also conduct comparative experiments to illustrate the effectiveness of mCVD. As Table 5 shows, "CADM-M" means the metric drop mCVD in the first epoch, using CVD instead. Experiment results indicate that CADM with mCVD in the first epoch can extremely improve the distance measurement and clustering performance. Therefore, mCVD can significantly address the negative impacts of center initialization randomness for CADM.

A.4. The Algorithm and Analysis of CADM

Theorem 1. *Proposed distance measurement is a distance metric.*

Proof. According to Eq. (3), it is obvious that the defined distances satisfy the following properties for any x_a, x_b , and $x_f \in S$.

- 1) $d(x_a, x_b) \geq 0$.
- 2) $x_a = x_b \Leftrightarrow d(x_a, x_b) = 0$.
- 3) $d(x_a, x_b) = d(x_b, x_a)$.
- 4) $d(x_a, x_b) \leq d(x_a, x_f) + d(x_b, x_f)$.

Therefore, the defined distance measures satisfy all the properties of a distance metric. ■

Table 4. Experiments with competitive distance metric in mixed datasets. "—" indicates that the algorithm is inapplicable or has not converged in one dataset.

Index	Dataset	HDM	GSM	LSM	CBDM	EBDM	UDM	HARR	COF	QGRL	CADM
CA	AA	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	<u>0.636 ± 0.01</u>	0.661 ± 0.03
	HF	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	<u>0.713 ± 0.03</u>	0.736 ± 0.03
	HD	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	<u>0.432 ± 0.02</u>	0.471 ± 0.03
	MM	0.818 ± 0.00	0.820 ± 0.00	0.831 ± 0.00	0.828 ± 0.00	0.807 ± 0.00	0.837 ± 0.00	0.818 ± 0.00	0.826 ± 0.01	0.830 ± 0.00	<u>0.832 ± 0.00</u>
Rank:		7.6	7.2	5.0	5.0	6.5	5.3	5.4	5.5	<u>2.5</u>	1.3
ARI	AA	0.002 ± 0.01	−0.002 ± 0.01	0.000 ± 0.02	—	0.025 ± 0.02	0.009 ± 0.02	0.005 ± 0.05	0.003 ± 0.05	<u>0.007 ± 0.02</u>	0.027 ± 0.04
	HF	0.009 ± 0.02	0.004 ± 0.02	0.001 ± 0.00	—	<u>0.124 ± 0.02</u>	0.044 ± 0.05	0.078 ± 0.00	0.018 ± 0.05	0.103 ± 0.04	0.140 ± 0.01
	HD	0.112 ± 0.03	0.154 ± 0.02	0.005 ± 0.02	—	0.062 ± 0.03	0.155 ± 0.04	<u>0.157 ± 0.05</u>	0.113 ± 0.06	0.147 ± 0.01	0.208 ± 0.03
	MM	0.325 ± 0.01	0.323 ± 0.00	0.345 ± 0.00	0.341 ± 0.00	0.305 ± 0.00	0.359 ± 0.00	0.379 ± 0.00	0.425 ± 0.00	0.434 ± 0.01	<u>0.429 ± 0.00</u>
Rank:		7.3	7.8	8.2	6.0	5.5	4.0	3.8	5.3	<u>3.3</u>	1.3
NMI	AA	0.003 ± 0.02	−0.002 ± 0.04	0.000 ± 0.00	—	<u>0.025 ± 0.03</u>	0.009 ± 0.03	0.009 ± 0.01	0.003 ± 0.04	0.002 ± 0.01	0.028 ± 0.03
	HF	0.001 ± 0.01	0.000 ± 0.01	0.000 ± 0.00	—	0.012 ± 0.03	0.046 ± 0.02	0.108 ± 0.02	0.003 ± 0.06	<u>0.109 ± 0.03</u>	0.124 ± 0.02
	HD	0.178 ± 0.03	0.139 ± 0.02	0.004 ± 0.01	—	0.133 ± 0.03	0.191 ± 0.04	0.209 ± 0.03	0.159 ± 0.02	<u>0.168 ± 0.02</u>	<u>0.198 ± 0.06</u>
	MM	0.325 ± 0.01	0.323 ± 0.00	0.345 ± 0.00	0.341 ± 0.00	0.305 ± 0.00	<u>0.359 ± 0.00</u>	0.327 ± 0.00	0.346 ± 0.00	0.349 ± 0.00	0.362 ± 0.00
Rank:		6.0	7.8	8.8	6.0	6.3	<u>3.1</u>	3.6	5.8	4.0	1.3
Total Rank:		7.0	7.6	7.3	6.0	6.1	4.2	4.3	5.5	<u>3.3</u>	1.3

Table 5. The ablation study results of CVD and mCVD

Index	Dataset	CADM-M	CADM
CA	NS	<u>0.257</u>	0.429
	LS	<u>0.541</u>	0.608
	VT	<u>0.879</u>	0.880
	MM	<u>0.782</u>	0.832
ARI	NS	<u>0.041</u>	0.085
	LS	<u>0.028</u>	0.314
	VT	0.573	0.615
	MM	<u>0.317</u>	0.429
NMI	NS	<u>0.041</u>	0.084
	LS	<u>0.186</u>	0.359
	VT	<u>0.470</u>	0.583
	MM	<u>0.245</u>	0.362
Total Rank:		<u>2</u>	1

Theorem 2. Distance metric defined can be extended to use in mixed data.

For the numerical attributes, due to their distance that is well-defined in real-number space, we directly calculate their distance using L1 norm and additionally utilize a simple "Add one" strategy, which means "Add one" after L1 norm, aligning the distance space between categorical attributes and numerical attributes. This strategy would not influence final clustering results. Similar to distance of categorical attributes, the distance of numerical attributes is zero if they are equal.

Therefore, the distance measurement for numerical attributes is defined as:

$$d_u(x_i^r, c_i^r) = \|x_i^r - c_i^r\| + 1, \quad (1)$$

where $A^r \in A^{num}$ and $\|\cdot\|$ is L1 norm.

Proof. Assume any categorical data x_a and $x_b \in S$. According to Eqs. (4) and (7) (in our article), it is obvious that $CVI^l(\cdot) \in [0, 1]$, and $\frac{1}{CVI^l(\cdot)} \in [1, +\infty]$. Therefore, CVD satisfies $d_a^l(x_a^r, x_b^r) \in [1, +\infty]$. Moreover, the range of the weight of attribute contributions satisfies $d_I(\cdot) \in [0, 1]$. As

aforementioned, their distance would be zero if $x_a^r = x_b^r$. Thus, the distance measurement of categorical attributes satisfies $d_c(x_a^r, x_b^r) \in \{0, [1, +\infty]\}$.

Furthermore, based on the definition in Eq. (1), the distance of any numerical data x_c and $x_d \in R$, satisfies: $d_u(x_c^r, x_d^r) \in \{0, [1, +\infty]\}$. Therefore, the distance space of the two types of attributes is the same (i.e., $\{0, [1, +\infty]\}$). Thus, modified CADM can be extended to mixed data. ■

Time complexity. The time complexity to update the CAI measure is $O(I \cdot k \cdot n \cdot d)$, where I is the number of iterations, k is the number of clusters, n is the dataset size and d is the data dimension. For updating the CVD function, the time complexity is $O(I \cdot k \cdot n \cdot d \cdot a)$, where a is the maximum size of the attribute's possible intra-values. Thus, the total time complexity is $O(I \cdot k \cdot n \cdot d \cdot a + I \cdot k \cdot n \cdot d)$.

Space complexity. The CAI measure is the size of $k \cdot d$, and the CVD metric is the size of $k \cdot d \cdot a$, therefore, the total space complexity is $O(k \cdot d \cdot a + k \cdot d)$.

The complexity should be efficient in most cases because I and k commonly are much smaller than n and d , and the size of a is normally small as well. Compared with other distance metrics (Zhang & Cheung, 2022, 2021), it will not cost more time and resources.

A.5. More Analysis on US dataset

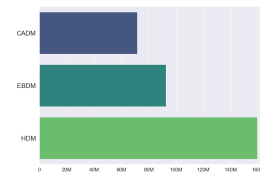


Fig. 1. Compared with efficient methods in the US dataset

Furthermore, although HDM and EBDM are faster than CADM, the Figure 1 indicates CADM has significant superiority in clustering performance by leveraging the within-cluster sum of squares (WSS) indicator, which calculates the cumulative distance between objects and their assigned cluster center without knowing the true labels. This means CADM is the best solution while considering both efficiency and accuracy on the tremendously large dataset, which demonstrates the wide usage range of the proposed distance metric.

A.6. Ablation Study of CADM

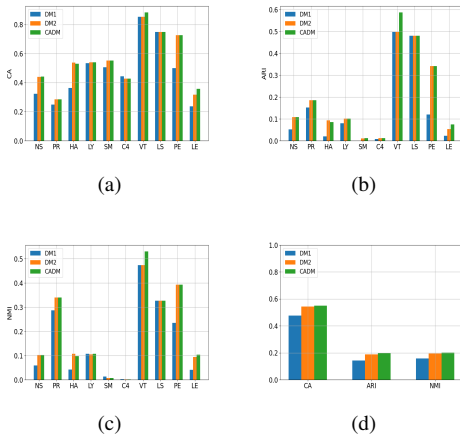


Fig. 2. Ablation study of clustering performance in the CA, ARI, and NMI indicators among ten categorical datasets. The last figure indicates the average performance of the ablated model compared with CADM.

In our ablation studies, the DM1 is the simple distance measurement using order information without the cluster-customized mechanism. DM2 is the cluster-customized adaptive measure based on the CVD that considers the distance difference between two attribute values in various clusters. The CADM adds the measurement of attribute importance to DM2. We conduct our ablation studies in ten categorical datasets and four mixed datasets receptively.

A.6.1. Ablation Study in Categorical Datasets

Based on our ablation study results in ten categorical datasets, which are shown in Figure 2, our observations are as follows:

The ablation study indicates that the effectiveness of CVD, as DM2 outperforms DM1 in almost all datasets among three indicators, showing the advantages of our designed cluster-customized mechanism. CADM is further superior to DM2, which illustrates the effectiveness of the CAI and the whole framework. However, due to CAI is designed to achieve minute adjustments for distance measurement,

therefore, the improvement would not be significant in most datasets.

Specifically, on 5 more complicated categorical datasets (i.e., NS, PR, HA, LY, SM), DM2 achieves tremendous improvement, which indicates that the distance difference of attribute values in various clusters is significant when there is a mixture of ordinal attributes and nominal attributes, so the superiority of the DM2 is obvious. Moreover, the performance of CADM is still superior to DM2, which proves that CAI can successfully weight the contributions of each attribute in complex conditions.

On pure ordinal datasets (i.e., C4, LE, LS), the experiment results of DM2 and CADM are not significantly superior to those of DM1 on LS and C4, because the number of LS attributes and the data size are small, making the advantages of DM2 less obvious. C4 is significantly more challenging than the others, with more attributes and a larger dataset size, which restricts drastic improvements in this dataset. On pure nominal datasets (i.e., VT, PE), the advantages of DM2 and CADM are obvious, which indicates that CVD and CAI can cooperatively provide reasonable measurements of nominal attribute values and nominal attribute category contributions, facilitating the entire distance metric accuracy and reliability. The Figure 2 (d) illustrates the average performance of DM1, DM2, and CADM in three indicators among ten datasets. The result straightforwardly indicates the effectiveness of CVD and CAI.

A.6.2. Ablation Study in Mixed Datasets

Based on our ablation study results of four mixed datasets in Figure 3, our observations are as follows: 1) CADM outperforms DM2 and DM1 in all datasets among the three indicators, which further illustrates the effectiveness of CAI. Especially, DM2 achieves extremely significant improvements compared with DM1, indicating that CVD can successfully project the distance space of categorical attributes into the distance space of numerical attributes, which is critical for heterogeneous data distance metric and clustering. 2) Compared with other datasets, the advantage of DM2 over DM1 on the HF dataset is not obvious. The reason is that HF has 7 numerical attributes and only 5 categorical attributes. This means the distance contributions of categorical attributes are smaller than numerical attributes. Although CADM can align the distance space between categorical and numerical attributes, it is still designed for categorical attributes, so when the distance contribution of categorical attributes is less, its effect will also be reduced, causing less improvement for clustering. 3) Figure 3 (d) shows the average performance of the ablated model in the mixed dataset. The results further demonstrate the necessity of the proposed mechanisms. As aforementioned, the contributions of categorical attributes are reduced, and the adjustment effect of CAI is weakened in most mixed datasets.

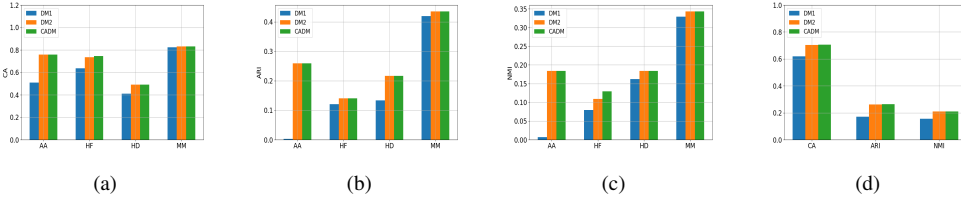


Fig. 3. Ablation study of clustering performance in the CA, ARI, and NMI indicators among four mixed datasets. The last figure indicates the average performance of the ablated model compared with CADM.

A.7. Visualization Analysis



Fig. 4. Distance measurement facilitates the clustering process. The left subfigure displays the data points with their original true labels and distances, and the right subfigure shows the data points with predicted labels calculated using the proposed distance.

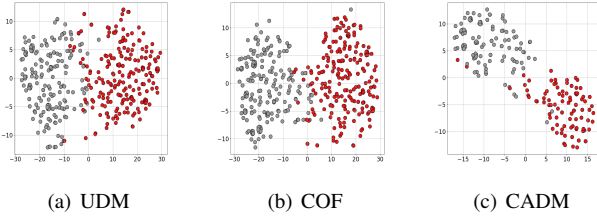


Fig. 5. t-SNE visualization of representations obtained by UDM, COF, and CADM on VT dataset. CADM achieves 88% accuracy in it. Red and gray mark the objects with the “true” cluster labels.

The Figure 4 indicates that our proposed distance measure can facilitate accurate clustering by making dissimilar points more separable.

Besides, Figure 5 shows the t-sne (Van der Maaten & Hinton, 2008) results of UDM (Zhang & Cheung, 2022), COF (Zhao et al., 2024), and CADM. It can be observed that true clusters in our CADM are more separable and distinguishable than COF and UDM, which indicates that CADM can make clusters more separable while maintaining each cluster gathering.

Moreover, we visualize the clustering results of mixed data by HARR (Zhang et al., 2025), QGRL (Chen et al., 2024), and CADM. HARR and QGRL are the SOTA models in the mixed dataset HF. It is noteworthy that HF is an imbalanced dataset, where cluster 1 has 203 data points, and cluster 2 has only 90 data points. From Figure 6, it can be observed that CADM produces more reliable clustering results, be-

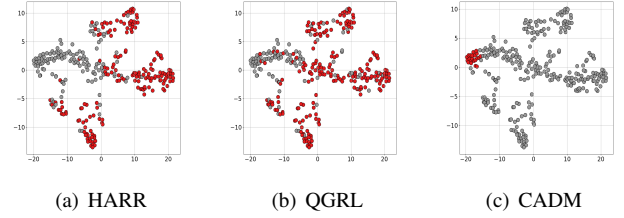


Fig. 6. t-SNE visualization of clustering results obtained by HARR, QGRL, and CADM on HF dataset. CADM achieves 73% accuracy in it. Red and gray mark the objects with the “true” cluster labels, which represent the two clusters of clustering results.

cause CVD can leverage the rival factor to reasonably zoom in and out of the distance between attribute values, which is useful when encountering the imbalanced data condition.

References

- 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*, 28(1):110–118, 2007.
- Pawel Bielski and Dustin Kottonau. Micro gas turbine electrical energy prediction. UCI Machine Learning Repository, 2024. <https://doi.org/10.24432/C58S4T>.
- Junyong 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*, pp. 297–306, 2024.
- Florian Espósito, Donato Malerba, V Tamma, HH Bock, et al. Classical resemblance measures. *Studies In Classification, Data Analysis, and Knowledge Organization*, 15: 139–152, 2000.
- David W Goodall. A new similarity index based on probability. *Biometrics*, pp. 882–907, 1966.

Dekang Lin et al. An information-theoretic definition of similarity. In *Proceedings of International Conference on Machine Learning*, volume 98, pp. 296–304, 1998.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008.

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*, volume 2, pp. 403–413. Elsevier Amsterdam, The Netherlands, 2005.

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*, 44(7):3560–3576, 2021.

Yiqun Zhang and Yiuming Cheung. A new distance metric exploiting heterogeneous interattribute relationship for ordinal-and-nominal-attribute data clustering. *IEEE Transactions on Cybernetics*, 52(2):758–771, 2022.

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*, 31(1):39–52, 2020. doi: 10.1109/TNNLS.2019.2899381.

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

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.