Robust Qualitative Data Clustering via Learnable Metric Space Fusion

Supplementary Material

APPENDIX

A. PROOFS OF THEOREM

Theorem 1. *The similarity measure* $\Gamma(\mathbf{x}_i, \mathbf{x}_j)$ *defined in the context of the MSF represents a valid distance metric.*

Proof. $\Gamma(\mathbf{x}_i, \mathbf{x}_j)$ follows non-negativity, symmetry, and triangle inequality for any $i \in \{1, 2, ..., N\}$ and $j \in \{1, 2, ..., N\}$.
Non-negativity: $\Gamma(\mathbf{x}_i, \mathbf{x}_j) \geq 0$. $\Gamma(\mathbf{x}_i, \mathbf{x}_j)$ is the fusion distance of each metric, which is always non-negative comprising non-negative weights;
Symmetry: $\Gamma(\mathbf{x}_i, \mathbf{x}_j) = \Gamma(\mathbf{x}_j, \mathbf{x}_i)$. Since the fusion metric is obtained by fusing base metrics, and these metrics inherently satisfy $\Gamma_m(\mathbf{x}_j, \mathbf{x}_i) = \Gamma_m(\mathbf{x}_i, \mathbf{x}_j)$, the fusion metric also obey the commutative law;
Triangle inequality: $\Gamma(\mathbf{x}_i, \mathbf{x}_j) \leq \Gamma(\mathbf{x}_i, \mathbf{x}_y) + \Gamma(\mathbf{x}_y, \mathbf{x}_j)$. In each base metric, the distance between two samples is unique. From sample $\mathbf{x}_i$ to $\mathbf{x}_j$, when detour through another sample $\mathbf{x}_y$, it necessarily involves extra distance from $\mathbf{x}_i$ to $\mathbf{x}_y$ and from $\mathbf{x}_y$ to $\mathbf{x}_j$. □

Probabilistic regularization is a method used to address the heterogeneity of different metrics, which can enhance the clustering performance of the fusion metric without distorting the properties satisfied by the original distance measures.

Theorem 2. *Given the distance matrices between possible values calculated by different metrics, the time complexity and space complexity of MSF is* $O(M\delta lNkd))$ *and* $O(MNk)$.

Proof. Assume δ_1 is the average number of iterations per weight update and δ_2 represents the additional iterations required for the convergence of **Q'**.

For each weight update, the distances between N samples and k modes across M metrics are calculated. Since the average number of iterations is δ_1, and the time complexity of searching the distance between two samples is $O(d)$, the time complexity for this phase is $O(M\delta_1 Nkd)$. 'd' represents the number of attributes. To calculate the distance between each pair of possible values, MSF only need to retrieve the corresponding positions in the distance matrix once. Therefore, calculating the distance between d possible values only requires d retrievals.

During the fusion phase, each update of **Q'** is accompanied by a re-evaluation of the distances between sample and modes, and the cluster probabilistic metric regularization strategy measures the probability values of N samples with k clusters across d attributes. Therefore, the time complexity for this phase is $O(M\delta_2 Nkd + MNkd)$.

In summary, assume that the weights are updated l times in total, the overall time complexity is $O(Ml(\delta_1 + \delta_2 + 1)Nkd)$. Therefore, the time complexity can be simplified to $O(M\delta lNkd)$, where δ represents the sum of constant terms δ_1 and $(\delta_2 + 1)$.

Furthermore, in terms of the required storage space for the algorithm, we only need to store the distances between all points and modes under each metric. Therefore, the time complexity can be summarized as $O(MNk)$. □

Since static data only needs to be calculated once for different metrics, in order to simplify the analysis process, the theorem does not analyze the complexity of calculating the similarity matrix between possible values based on the base metric. If readers need to know more, please read the original paper of the metric we use.

TABLE S.1
STATISTICS OF THE 10 DATASETS. d, n, AND k^* REPRESENT THE NUMBERS OF ATTRIBUTES, SAMPLES, AND TRUE CLUSTERS, RESPECTIVELY.

No.	Dataset	Abbrev.	d	n	k^*
1	Tic-Tac-Toe	TT	9	958	2
2	Caesarian Section	CS	4	80	2
3	Car Evaluation	CE	6	1728	4
4	Congressional Voting	VT	16	435	2
5	Lymphography	LG	18	148	4
6	Australia Credit	AC	8	690	2
7	Employee Rejection	ER	4	1000	9
8	Dermatology	DT	33	366	6
9	Assistant Evaluation	AE	4	72	3
10	Lecturer Evaluation	LE	4	1000	5

B. DETAILS OF THE EXPERIMENT

B.1. Details of the dataset

Ten datasets from various fields are obtained from the UCI machine learning repository [1] for the experiments, and the statistical informations is show in Table S.1. All the datasets are pre-processed by removing the samples with missing values [2], [3].

TABLE S.2
CLUSTERING PERFORMANCE EVALUATED BY CA. ORANGE AND GRAY REPRESENT THE BEST AND SECOND-BEST RESULT. THE SYMBOL "•" INDICATES THAT THE MSF METHOD AND OTHER COMPARATIVE METHODS HAVE PASSED THE 95% CONFIDENCE TEST BASED ON THE WILCOXON SIGNED-RANK TEST.

Dataset	EH+MMR	ECPCS-HC	LWEA	ELSC	LROLML-2	SCPP-a	H2H	COForest	MSF(ours)
TT	0.5294±0.03	0.6162±0.00	0.6204±0.02	0.6208±0.00	0.6406±0.04	0.6526±0.00	0.5433±0.00	0.5731±0.00	0.6816±0.00 •
CS	0.5925±0.06	0.5150±0.01	0.5138±0.01	0.6013±0.00	0.6038±0.04	0.5863±0.04	0.6175±0.02	0.6313±0.04	0.6500±0.00
CE	0.4108±0.03	0.5946±0.04	0.5937±0.05	0.3983±0.05	0.5155±0.12	0.6987±0.00	0.3955±0.11	0.3542±0.04	0.5883±0.00
VT	0.8713±0.00	0.8437±0.01	0.8600±0.03	0.8391±0.00	0.7224±0.07	0.8626±0.00	0.8736±0.00	0.8761±0.00	0.8782±0.00 •
LG	0.4730±0.04	0.4966±0.03	0.4956±0.03	0.4770±0.01	0.4628±0.08	0.5662±0.05	0.5250±0.04	0.5088±0.07	0.5608±0.00
AC	0.7942±0.00	0.5106±0.01	0.5330±0.05	0.5116±0.00	0.5509±0.01	0.6981±0.12	0.7703±0.08	0.7857±0.11	0.8087±0.00 •
ER	0.1986±0.01	0.1943±0.01	0.1917±0.01	0.1723±0.01	0.2010±0.01	0.1948±0.01	0.1836±0.01	0.1929±0.01	0.1920±0.00
DT	0.7492±0.06	0.5492±0.08	0.5455±0.06	0.7472±0.00	0.5075±0.08	0.0302±0.09	0.7800±0.04	0.7048±0.08	0.7877±0.00
AE	0.5681±0.06	0.6000±0.04	0.5889±0.04	0.5042±0.05	0.5278±0.05	0.4292±0.07	0.5181±0.07	0.5583±0.09	0.6944±0.00 •
LE	0.3116±0.02	0.3614±0.02	0.3504±0.02	0.3014±0.00	0.3526±0.03	0.4043±0.00	0.3144±0.02	0.3155±0.2	0.3240±0.00

B.2. Performance comparison of single metric and multi-metric fusion

We have conducted an intuitive comparison of the clustering performance of single metric and multi-metric fusion, as shown in Figure S.1. For completeness, comparisons on all datasets are provided here. 'RW' refers to a method that assigns weights to individual metrics randomly based on MSF, while 'w/o W' assigns equal weights to each metric based on MSF. UDM, CBDM, and EBDM represent clustering methods using K-modes with three different metrics. It can be observed that the performance of 'w/o W' is higher than the single metric on almost all datasets, which validates the effectiveness of the fusion method. On seven out of ten datasets, 'RW' performs better than 'w/o W' method, which further hints that learning a set of weights for specific clustering tasks in metric fusion is likely to achieve more robust and accurate clustering performance.

B.3. Clustering Performance

Clustering performance of different methods and significance test are compared in Table S.2 and Table S.3. The bast and second-bast results are marked in orange and grey on each dataset, respectively. Significance test is conducted between the best and the second-best results by Wilcoxon signed rank test [4] with a 95% confidence interval, and significant difference is indicated by the symbol "•". **EH+MMR** utilizes the ensemble process of the ECPCS-HC method and employs the results obtained from three base metrics as the base results for the clustering ensemble process. The observations include the following four aspects: (1) Overall, MSF performs the best or second-best on almost all dataset. (2) MSF achieves the best performance in 6 out of 10 datasets based on CA, while it achieved the second-best result in the LG dataset. (3) The MSF method achieves the best performance in 5 out of 10 datasets based on NMI, while it achieved the second-best result in the CE, LG, AC datasets. (4) Although MSF does not have the best performance on CE, ER and LE datasets, it is not surpassed by much by the winners and still very competitive.

B.4. Efficiency evaluation

To evaluate the efficiency of MSF, large synthetic datasets are randomly generated with different scales of attributes and samples. Specifically, we generate by: 1) Fixing the number of attributes at $l = 20$ and increasing the number of samples n from 10k to 100k with step-size 10k, and 2) Fixing sample size at $n = 2k$ and increasing the number of attributes l from 1k to 10k with step-size 1k, where 'k' indicates 'kilo'.

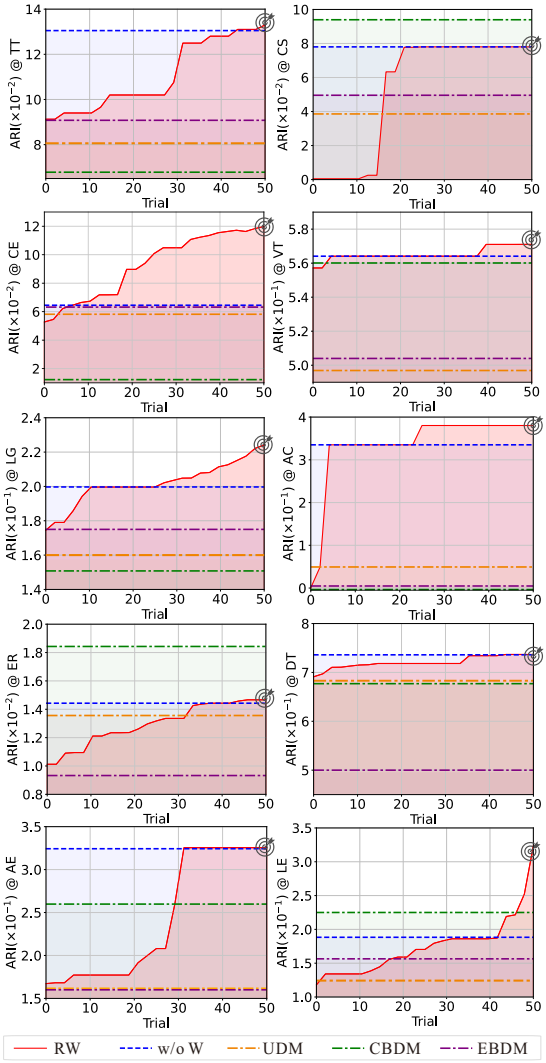

Fig. S.1. Supplementary material about toy-example in Figure 1: performance comparison of all datasets

TABLE S.3
CLUSTERING PERFORMANCE EVALUATED BY NMI. ORANGE AND GRAY REPRESENT THE BEST AND SECOND-BEST RESULT. THE SYMBOL " • " INDICATES THAT THE MSF METHOD AND OTHER COMPARATIVE METHODS HAVE PASSED THE 95% CONFIDENCE TEST BASED ON THE WILCOXON SIGNED-RANK TEST.

Dataset	EH+MMR	ECPCS-HC	LWEA	ELSC	LROLML-2	SCPP-a	H2H	COForest	MSF(ours)
TT	0.0071±0.00	0.0071±0.00	0.0070±0.10	0.0420±0.00	0.0041±0.01	0.0016±0.00	0.0110±0.01	0.0106±0.01	0.1052±0.00 •
CS	0.0536±0.04	0.0204±0.00	0.0265±0.01	0.0287±0.00	0.0475±0.03	0.0780±0.04	0.0448±0.02	0.0749±0.03	0.0975±0.00 •
CE	0.1731±0.05	0.0402±0.01	0.0407±0.02	0.0349±0.01	0.0529±0.03	0.0046±0.01	0.0715±0.05	0.1132±0.07	0.1163±0.00
VT	0.4901±0.01	0.4073±0.04	0.4355±0.06	0.4409±0.00	0.1667±0.11	0.4515±0.00	0.4893±0.00	0.4897±0.00	0.4993±0.00
LG	0.1328±0.04	0.1546±0.02	0.1701±0.03	0.2392±0.01	0.1088±0.04	0.1594±0.07	0.2172±0.02	0.1811±0.05	0.2311±0.00
AC	0.2668±0.00	0.0451±0.00	0.0537±0.04	0.0487±0.00	0.0023±0.00	0.1615±0.13	0.2402±0.08	0.3042±0.17	0.3027±0.00
ER	0.0667±0.00	0.0708±0.00	0.0680±0.00	0.0355±0.00	0.0614±0.01	0.0411±0.02	0.0494±0.00	0.0049±0.01	0.0533±0.00
DT	0.7948±0.05	0.6367±0.04	0.6196±0.03	0.7171±0.00	0.4501±0.10	0.0020±0.01	0.8023±0.02	0.8089±0.06	0.8567±0.00 •
AE	0.1934±0.05	0.2076±0.07	0.2112±0.05	0.1892±0.05	0.2312±0.07	0.0816±0.09	0.2249±0.08	0.2340±0.08	0.3795±0.00 •
LE	0.0551±0.01	0.0467±0.02	0.0319±0.01	0.0375±0.00	0.0638±0.01	0.0080±0.00	0.0574±0.02	0.0560±0.02	0.0573±0.00

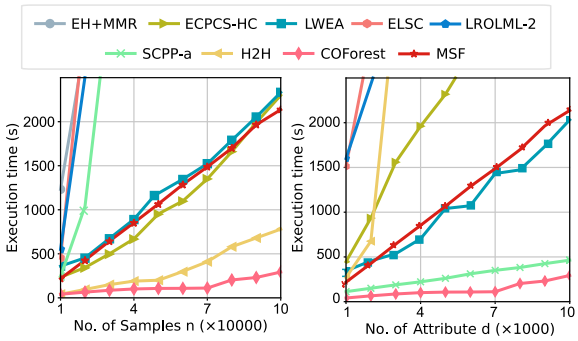

Fig. S.2. Execution time on synthetic datasets under increasing scales of samples and attributes

Note that each attribute has five possible values, and the number of clusters k is consistently set to five. To simplify the computational process, we use the Hamming distance as the base metric for the fusion framework. The execution time of the 9 methods is demonstrated in Figure S.2. It can be seen that the execution time of MSF is lower than or similar with almost all methods. Moreover, the increasing trend of the execution time of MSF is almost linear with n, which is consistent with the time complexity analysis of Theorem 2. In summary, MSF is efficient compared to the state-of-the-art methods and does not incur too much additional computational cost compared to the simplest methods.

REFERENCES

[1] M. Kelly, R. Longjohn, and K. Nottingham, "UCI machine learning repository."

[2] J. Chen, Y. Ji, R. Zou, Y. Zhang, and Y.-m. 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.

[3] S. Cai, Y. Zhang, X. Luo, Y.-M. Cheung, H. Jia, and P. Liu, "Robust categorical data clustering guided by multi-granular competitive learning," in *2024 IEEE 44th International Conference on Distributed Computing Systems (ICDCS)*, 2024, pp. 288–299.

[4] S. Sidney, "Nonparametric statistics for the behavioral sciences," *Journal of Nervous and Mental Disease*, vol. 125, no. 3, p. 497, 1957.