ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10889806

Robust Qualitative Data Clustering via Learnable Multi-Metric Space Fusion

Sen Feng[1], Mingjie Zhao[1], Zhanpei Huang[1], Yuzhu Ji[1], Yiqun Zhang[1,2,*], Yiu-ming Cheung[2]
[1]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China
[2]Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China

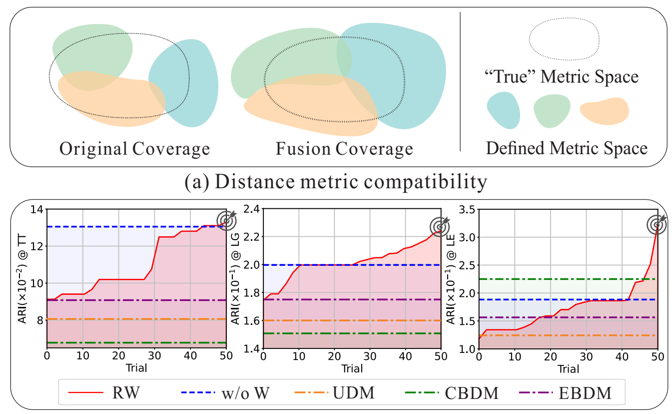
***Abstract*—Understanding categorical data with vague qualitative values by forming clusters is crucial in many data-driven AI fields. Compared with numerical data with its quantitative values embedded in well-defined Euclidean distance space, distances of the qualitative values are naturally unknown and are specially defined for certain data types or tasks. This paper, therefore, proposes a distance metric space fusion framework, which learns to fuse multiple distance metrics to form a statistical information-complete and prior knowledge-comprehensive metric for robust and accurate cluster analysis of qualitative data. To better serve various clustering tasks, the metric fusion objective is incorporated into the clustering objective through iterative learning. It turns out that the proposed method stably demonstrates superiority on various challenging real benchmark datasets. Extensive experiments including significance tests, ablation studies, etc. validate its efficacy. Source code of the proposed method is available at https://github.com/Sen-Feng/ICASSP-MSF/tree/main/CODE.**

***Index Terms*—Cluster analysis, categorical data, unsupervised learning, metric space learning, robust and accurate clustering**

I. Introduction

Cluster analysis is crucial in many practical applications [1]–[4], such as speech recognition, medical data analysis, etc. As a common data type, categorical data [5]–[8] composed of qualitative valued attributes are hard to embed into the well-defined Euclidean distance space appropriately. Hamming distance [9] is a conventional distance metric for categorical data, which simply measures distance based on whether two values are identical. More advanced distance metrics have also been proposed to define the distances based on the: (1) Information extracted from correlated attributes [10]–[13], and (2) Information entropy computed according to the occurrence probabilities of attribute values [14]–[16]. Since the existing distance metrics are defined from different perspectives, they can only well-serve a certain type of dataset, which we call it incompleteness of distance metric.

In the literature, cluster ensemble and distance metric learning methods have been proposed to relieve the influence of metric incompleteness to a certain extent. A common ensemble



(a) Distance metric compatibility

(b) Performance comparison of single metric and multi-metric fusion

Fig. 1. Original metrics vs. multi-metric fusion. (a) intuitively demonstrates the coverage of the true metric space by the original metrics and multi-metric fusion. (b) illustrates the effect shown in (a) through demo experiments, where the clustering accuracy in terms of the Adjusted Rand Index (ARI) [17] of three representative distance metrics (i.e., UDM [12], CBDM [11], and EBDM [14]) and their fusions formed through Random Weighting (RW) and without Weighting (w/o W) are compared. The fusion is performed in the sample partition stage in the clustering. It can be observed that even the simple unweighted fusion (i.e., ‘w/o W’, which is equivalent to metric fusion with equal weights) can easily surpass each individual metric on the TT and LG datasets. Moreover, fusion with random weights can further obviously outperform ‘w/o W’ on LG and LE datasets. Therefore, our goal is to learn a fusion of multiple metrics that can reach the ‘target’ performance in (b).

way is to combine the sample-cluster affiliation obtained by different clustering approaches [18]–[20]. Some approaches [21]–[24] also attempt to generate multiple clustering results by using a single clustering algorithm with different parameter settings. These approaches are mainly designed for an algorithm-level complement. Although clustering results corresponding to different distance metrics can also be combined for ensemble clustering (see the performance of the combination of the conventional metrics, i,e., EH+MMR in Table II), its effectiveness is very limited due to the possible homogeneity and contradiction of the adopted distance metrics.

Distance metric learning usually learns a distance structure that is more suitable for certain clustering tasks. Graph-based learning approaches [25]–[30] learn the relationship among possible values based on their graph structures constructed according to prior knowledge of data. To relieve the prior knowledge limitation, the work proposed in [31], [32] further extends the possible value graphs into minimal spanning trees for more concise and flexible metric learning. On the other hand,

* Corresponding author: Yiqun Zhang (yqzhang@gdut.edu.cn).

This work was supported in part by the National Natural Science Foundation of China (NSFC) under grants: 62476063, 62102097, and 62302104, the NSFC/Research Grants Council (RGC) Joint Research Scheme under grant: N_HKBU214/21, the Natural Science Foundation of Guangdong Province under grants: 2024A1313010039, 2023A1515012855, and 2023A1515012884, the General Research Fund of RGC under grants: 12201321, 12202622, and 12201323, the RGC Senior Research Fellow Scheme under grant: SRFS2324-2S02, and the Science and Technology Planning Project of Guangzhou under grant SL2023A04J01625.

probability-based approaches [33]–[35] are also popular. They model sample-cluster affiliation according to the occurrence probability of the sample values in clusters, and combined with certain clustering algorithms [36]–[38] for clustering. All the above-mentioned methods achieve considerable improvements in clustering accuracy. However, their effectiveness is highly sensitive to the unforeseeable compatibility of their adopted prior knowledge to certain clustering tasks.

To sum up, almost all the existing qualitative data clustering solutions suffer from the limitations brought by the adopted distance metrics. That is, metric learning methods overly rely on prior knowledge whilst ensemble methods lack a learning mechanism to coordinate multiple metrics or partitions. As intuitively visualized in Figure 1 (a), a single distance metric is difficult to precisely 'hit' the true distance metric of a specific clustering task. Multiple original metrics also difficult to fully cover the target metric space without tuning. We illustrate this through a preliminary experiment shown in Figure 1 (b), which hints that learning a set of weights w.r.t. certain clustering tasks for metric fusion is promising for categorical data clustering.

This paper, therefore, proposes a Multi-metric Space Fusion (MSF) framework that learns to integrate various distance metrics for more robust and accurate clustering. Unlike traditional ensemble methods that typically merge the clustering results, our approach integrates multiple metrics for clustering and iteratively updates the integration and clustering to search for an optimal solution. More specifically, the importance of each metric is quantified as the likelihood of a metric in forming compact clusters. To address the heterogeneity of different metrics, a probabilistic regularization is introduced to map the sample-cluster similarity into a homogeneous space for the metric importance quantification. It turns out that the proposed MSF and the iterative learning algorithm bridge the gap between multiple distance metrics and certain clustering tasks, thus achieving robust and accurate clustering performance on qualitative datasets from various domains. Three main contributions of this paper are summarized as follows:

- A metric fusion clustering framework is proposed to learn the combination of metrics during clustering. Compared to the ensemble approaches, it can flexibly adapt to clustering tasks. In contrast to existing metric space learning, it compensates for the limitations of single metric.
- To address the fusion bias caused by the heterogeneity of metrics, we propose an entropy-based measure to quantify the importance of metrics, and also develop a probabilistic regularization strategy to align the importance.
- The proposed method demonstrates merits of robustness, interpretability, and scalability. In addition to its superior clustering accuracy, it also supports the fusion of user-specified metrics, and can be continuously and efficiently applied to fuse newly developed advanced metrics.

II. PROPOSED METHOD

Table I lists the symbols that appear frequently in this work, where we use bold to indicate a vector or matrix.

TABLE I
EXPLANATION OF SYMBOLS

Symbol	Explanation
$\mathbf{X}$	The whole dataset with n samples
N and k	Number of samples and clusters
$\mathbf{x}_i$ and $x_{i,r}$	i-th data sample and its r-th attribution value
$\mathbf{a}_r$ and $v_{r,s}$	r-th attribute and its s-th possible value
o_r	Number of possible values for the r attribute
$\mathbf{Q}$ and $q_{i,j}$	Membership matrix and its (i,j)-th entry
c_j and $\mathbf{u}_j$	j-th cluster and its mode
$\mathbf{\Upsilon}_m$ and γ_m^j	The result of m-th metric and its j-th cluster
w_m	m-th weight of metric in fusion metrics
E_m	Normalized entropy of m-th metric

The multi-metric fusion is realized by learning the importance of each single metric and fusing their measured sample-cluster affiliation for clustering. The metric importance learning and clustering are iteratively performed to reach a local optimum of the problem. Given M metrics for fusion, we use a weight vector $\mathbf{W} = [w_1, w_2, \dots, w_M]$ to indicate their importance. With the current sample partition $\mathbf{Q}$ and cluster modes $\mathbf{U} = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k\}$, we first use each single metric (e.g., the m-th metric) to perform k-modes [39] clustering upon fixed $\mathbf{U}$ until convergence and obtain the clustering results $\mathbf{\Upsilon}_m$ to ensure the cluster information sufficiency of the following metric importance quantification:

$$w_m = w_m - \lambda(E_m - avg(\mathbf{E})), \tag{1}$$

where λ is the learning rate, and the normalized entropy vector $\mathbf{E} = [E_1, E_2, \dots, E_M]$ reflects importance of metrics by

$$E_m = \sum_{j=1}^{k} \sum_{r=1}^{d} \frac{\sum_{s=1}^{o_r} p_{v_{r,s}|\gamma_m^j} log(p_{v_{r,s}|\gamma_m^j})}{\sum_{s=1}^{v_r} p_{v_{r,s}|\mathbf{X}} log(p_{v_{r,s}|\mathbf{X}})}, \tag{2}$$

with the conditional probabilities of a possible value $v_{r,s}$ within cluster γ_m^j computed as

$$p_{v_{r,s}|\gamma_m^j} = \frac{\Psi_{a_r=s}(\gamma_m^j)}{\Psi_{a_r \neq NULL}(\gamma_m^j)}, \tag{3}$$

where $\Psi_{a_r=s}(\gamma_m^j)$ counts the number of samples in cluster γ_m^j that have the value $v_{r,s}$ for attribute a_r, and $\Psi_{a_r \neq NULL}(\gamma_m^j)$ indicates the number of objects in cluster γ_m^j that take values from attribute a_r.

Remark 1 (Rationality of the normalized entropy and $(E_m - avg(\mathbf{E}))$ in Eq. (1))**.** *Based on the principle that a metric with better clustering performance is deemed more important, we quantify the performance using normalized entropy $\mathbf{E}$ to circumvent the heterogeneity of metrics, where a lower E_m indicates higher intra-cluster similarity. Accordingly, an important metric will be assigned with a higher weight, and thus we use $(E_m - avg(\mathbf{E}))$ in Eq. (1) to ensure that the summation of updated weights still equals 1 and also facilitate a connection among all the metrics.*

To facilitate an effective update of $\mathbf{W}$, we adopt the common Adam [40] optimizer in the following cluster learning process. With the updated importance of metrics $\mathbf{W}$, a new round of clustering is performed based on the fusion metric

$$\Gamma(\mathbf{x}_i, \mathbf{u}_j) = \sum_{m=1}^{M} w_m \Gamma_m(\mathbf{x}_i, \mathbf{u}_j) \tag{4}$$

to obtain more reasonable data sample partition, where $\Gamma_m(\mathbf{x}_i, \mathbf{u}_j)$ is the similarity between $\mathbf{x}_i$ and $\mathbf{u}_j$ measured by the m-th metric, and clustering objective can be formulated as

$$L = \sum_{i=1}^{n} \sum_{j=1}^{k} q_{i,j} \Gamma(\mathbf{x}_i, \mathbf{u}_j) \tag{5}$$

to maximize the intra-cluster similarity. Here, $\mathbf{Q}$ is an $N \times k$ matrix with its (i,j)-th entry $q_{i,j} \in \{0,1\}$ indicating the affiliation of sample $\mathbf{x}_i$ to cluster c_j by

$$q_{i,j} = \begin{cases} 1, & \text{if } j = \arg\max_{y} \Gamma(\mathbf{x}_i, \mathbf{u}_y) \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Remark 2 (Relationship between $\mathbf{E}$ and L)**.** *For the qualitative value of categorical data, using entropy to measure cluster quality is consistent with the objective L, and thus updating $\mathbf{W}$ can be viewed as a relaxed maximization process of L. More specifically, Eq. (1) enhances the contribution of metrics with a smaller entropy in $\mathbf{E}$, leading to an effective increase in L. When $\mathbf{E}$ stabilizes, the fusion metric space tends to be fixed, and the whole clustering algorithm converges.*

As for the definition of $\Gamma_m(\mathbf{x}_i, \mathbf{u}_j)$, we propose a novel cluster probabilistic metric regularization strategy to unify various metrics that may be heterogeneous in terms of value meaning and magnitude. For each metric Γ_m, we form a corresponding partition $\mathbf{Q}_m$ of $\mathbf{X}$ using the original metric based on the same current modes $\mathbf{U}$ to ensure consistency. Then the occurrence probabilities of the possible values within each cluster are utilized to map an original metric to a probabilistic space by

$$\Gamma_m(\mathbf{x}_i, \mathbf{u}_j) = \frac{1}{d} \sum_{r=1}^{d} \frac{\Psi_{a_r = x_{i,r}}(\gamma_m^j)}{\Psi_{a_r \neq NULL}(\gamma_m^j)}, \tag{7}$$

which represents the probability that $\mathbf{x}_i$ belongs to γ_m^j indicated by the m-th metric.

Remark 3 (Necessity of the cluster probabilistic metric regularization)**.** *It unifies the heterogeneous metrics that may simultaneously include similarity metric and distance metric in different value magnitude. Extracting information from their own partitions $\mathbf{Q}_m$ also preserves their metric characteristics.*

The whole algorithm is summarized as Algorithm 1. According to our analyses, the similarity measure $\Gamma(\mathbf{x}_i, \mathbf{x}_j)$ defined in the context of the MSF represents a valid distance metric. Given the distance matrices between possible values calculated by different metrics, the time complexity and space complexity of MSF are $O(M\delta lNkd)$ and $O(MNk)$, respectively, which indicate its lightweight and efficiency.

Algorithm 1: MSF: Multi-Metric Space Fusion

Input: Dataset $\mathbf{X}$, No. of clusters k, No. of metrics M
Output: Fusion metric $\Gamma(\cdot,\cdot)$, Partition $\mathbf{Q}$

Step 1: Set the timestamp by $\tau = 0$; Initialize $\mathbf{W}^\tau$ by $w_m = \frac{1}{M}$; Initialize $\mathbf{U}^\tau$; Set $\mathbf{Q}^\tau$ and $\mathbf{Q}'$ to $\mathbf{0}$;

Step 2: Fix $\mathbf{U}^\tau$, iteratively update $\mathbf{\Upsilon}_m^\tau$ for each base metric until $\mathbf{\Upsilon}_m^\tau$ remain unchanged. Then we obtain $\mathbf{\Upsilon}_m^{\tau+1}$, and update $\mathbf{W}^{\tau+1}$;

Step 3: Fix $\mathbf{W}^{\tau+1}$, iteratively update $\mathbf{Q}'$ until all the values of $\mathbf{Q}'$ remain unchanged. Then update $\mathbf{U}^{\tau+1}$ by $\mathbf{Q}'$. If $\mathbf{Q}^\tau \neq \mathbf{Q}'$, set $\mathbf{Q}^{\tau+1} = \mathbf{Q}'$, $\tau = \tau + 1$, and go to **Step 2**; Otherwise, stop, Output $\Gamma(\cdot,\cdot)$ and $\mathbf{Q}^\tau$.

III. Experiments

Experimental settings are briefly described below. (1) **Three base metrics** are chosen for fusion by the proposed MSF, namely UDM [12], CBDM [11], and EBDM [14]. (2) **Ten datasets** from various fields are utilized for the experiments. Their detailed descriptions and statistics can be found by searching "Tic-Tac-Toe (TT)", "Caesarian Section (CS)", "Car Evaluation (CE)", "Congressional Voting (VT)", "Lymphography (LG)", "Australia Credit (AC)", "Employee Rejection (ER)", "Dermatology (DT)", "Assistant Evaluation (AE)", and "Lecturer Evaluation (LE)" in the website of the UCI machine learning repository [41]. (3) **Nine comparison methods**, including six clustering ensemble methods ECPCS-HC [20], EH+MMR [1], LWEA [19], ELSC [22], LROLML-2 [23], and SCPP-a [21]; two state-of-the-art metric space learning methods H2H [26], and COForest [32]; and a baseline clustering method k-modes (KM) [39]. In addition, six ablation variants of MSF are formed, including three MSF variants that only adopting a single metric without fusion learning, and three MSF variants that ablate one of or both of the core components, i.e., Weight Learning (WL) strategy, Metric Regularization (MR) strategy. (4) **Evaluation index**, i.e., Adjusted Rand Index (ARI) [17], [42], [43] with value range $[-1, 1]$ is adopted for evaluating clustering performance. We also conducted the Wilcoxon signed rank test [44] with 95% confidence interval between the best and second-best results, and a significant difference is indicated by symbol "•". All the experiments are coded using Python 3.11.

Clustering performance and significance test results are shown in Table II. The observations include the following three aspects: (1) Overall, MSF performs very stable and remains the best on all datasets. On seven out of ten datasets, MSF performs significantly better than the runner-up method, indicating the superiority of MSF. (2) Compared with the former six clustering ensemble methods, the superiority of MSF is more obvious, as multiple metrics can complement each other and their fusion is learned by MSF. (3) MSF does not significantly outperform the metric learning counterparts

[1] EH+MMR utilizes the ensemble strategy of ECPCS-HC (EH) to combine Multi-Metric clustering Results (MMR) generated by k-modes clustering algorithm using each of the three metrics adopted by our proposed MSF.

TABLE II

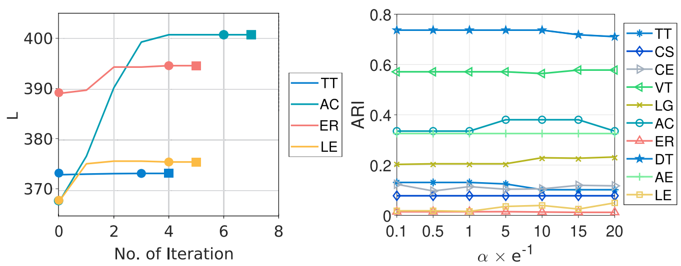
CLUSTERING PERFORMANCE W.R.T. ARI. ORANGE AND GRAY REPRESENT THE BEST AND SECOND-BEST RESULTS, RESPECTIVELY. THE SYMBOL "●" INDICATES THAT OUR MSF PASSES THE WILCOXON SIGNED-RANK TEST UNDER 95% CONFIDENCE INTERVAL AGAINST THE RUNNER-UP COUNTERPART.

Dataset	KM	EH+MMR	ECPCS-HC	LWEA	ELSC	LROLML-2	SCPP-a	H2H	COForest	MSF (ours)
TT	0.0166±0.00	0.0034±0.01	-0.0189±0.00	-0.0097±0.02	0.0572±0.00	0.0004±0.01	-0.0007±0.00	0.0077±0.01	0.0219±0.02	0.1310±0.00 ●
CS	0.0111±0.00	0.0368±0.04	-0.0174±0.00	-0.0173±0.00	0.0290±0.00	0.0362±0.03	0.0192±0.03	0.0443±0.02	0.0628±0.03	0.0780±0.00
CE	0.0304±0.00	0.0222±0.07	0.0293±0.05	0.0492±0.08	0.0499±0.01	-0.0197±0.04	-0.0006±0.00	0.0010±0.04	0.0444±0.05	0.1237±0.00 ●
VT	0.5194±0.00	0.5504±0.01	0.4720±0.04	0.5198±0.07	-0.0170±0.00	0.1989±0.13	0.5253±0.00	0.5572±0.00	0.5647±0.00	0.5710±0.00 ●
LG	0.0943±0.00	0.0970±0.04	0.1097±0.02	0.1343±0.04	0.1782±0.01	0.0620±0.03	0.0488±0.03	0.1957±0.03	0.1395±0.06	0.2261±0.00 ●
AC	0.2549±0.01	0.3453±0.00	-0.0081±0.00	0.0064±0.03	-0.0083±0.00	0.0034±0.00	0.2073±0.17	0.3107±0.11	0.2367±0.19	0.3803±0.00 ●
ER	0.0104±0.00	0.0127±0.00	0.0114±0.00	0.0082±0.00	0.0069±0.00	0.0129±0.00	0.0056±0.01	0.0077±0.00	0.0123±0.00	0.0148±0.00 ●
DT	0.4677±0.01	0.6794±0.09	0.4029±0.09	0.3747±0.06	0.5863±0.00	0.3093±0.13	-0.0006±0.00	0.7364±0.07	0.6605±0.11	0.7367±0.00
AE	0.1227±0.00	0.1439±0.07	0.1691±0.04	0.1491±0.06	0.1149±0.06	0.1731±0.07	0.0241±0.09	0.1729±0.08	0.1640±0.09	0.3255±0.00 ●
LE	0.0254±0.00	0.0283±0.01	0.0277±0.02	0.0197±0.01	0.0149±0.00	0.0105±0.02	0.0008±0.00	0.0325±0.01	0.0346±0.2	0.0355±0.00

TABLE III

CLUSTERING PERFORMANCE W.R.T. ARI OF ABLATED MSF VERSIONS.

Dataset	MSF with metric:			MSF without module:			MSF
	UDM	CBDM	EBDM	WL + MR	WL	MR	
TT	0.0806	0.0678	0.0908	0.0961	0.1305	0.0961	0.1310
CS	0.0386	0.0939	0.0496	0.0498	0.0780	0.0498	0.0780
CE	0.0582	0.0123	0.0602	0.0264	0.0657	0.0246	0.1237
VT	0.4969	0.5601	0.5034	0.5300	0.5641	0.5232	0.5710
LG	0.1575	0.1508	0.1747	0.0950	0.1997	0.1015	0.2261
AC	0.0493	-0.0035	0.0048	0.0048	0.3351	0.0048	0.3803
ER	0.0136	0.0184	0.0093	0.0122	0.0144	0.0122	0.0148
DT	0.6827	0.6770	0.4964	0.6588	0.7367	0.6588	0.7367
AE	0.1600	0.2590	0.1600	0.2052	0.3255	0.2353	0.3255
LE	0.0124	0.0225	0.0156	0.0184	0.0183	0.0201	0.0355
Average	0.1750	0.2038	0.1565	0.1697	0.2467	0.1727	0.2615



(a) Convergence curves (b) Learning rate curves

Fig. 2. (a) Convergence curves of MSF, dots and squares represent the iterations of metric weights updating and convergence, respectively; (b) ARI performance of MSF under different learning rates.

on the CS, DT, and LE datasets. The reason could be that CS and DT are both small-scale datasets. Insufficient statistical information cannot well support the metric learning and our fusion learning, thus causing similar-level performance. As for the LE dataset, since it is composed of only ordinal valued attributes with a relatively simple distance structure, there is not much room for improvements by the metric learning and our fusion learning.

Ablation studies of MSF from the perspectives of adopted metrics and key technical components are illustrated in the left and right parts of Table III, respectively. Observations for the metric aspect ablation are two-fold: (1) MSF outperforms all its single metric versions on most datasets. This indicates that MSF can effectively complement multiple metrics to reach higher clustering accuracy; (2) Performance of MSF with different single metrics varies on different datasets. This indicates that different metrics have their own bias for datasets, which again proves the necessity of metric fusion. Observations for ablating the key components are two-fold: (1) MSF performs better on most datasets. This generally validates the effectiveness of our MSF framework design; (2) The performance of MSF without both WL and MR is worse than MSF without only one of the two key components. This proves the effectiveness of WL and MR, respectively.

Converge and learning rate are evaluated in Figure 2 (a) and (b), respectively. The vertical axis in Figure 2 (a) represents the value of objective function L. The square markers indicate the iterations at which MSF converges, and the circle markers indicate the iterations of weights updating. It can be observed that L experiences a monotonic increase on different datasets and converges quickly within seven iterations, which intuitively illustrates the learning effectiveness and efficiency of MSF. It can be seen from Figure 2 (b) that the clustering performance of MSF is very stable across different datasets under a wide range of learning rates. This indicates that the learning rate is easy to set and we suggest a setting of $\alpha = [0.05, 0.1]$. In our released code, we uniformly set $\alpha = 0.1$, which can obtain satisfactory clustering performance.

IV. CONCLUDING REMARKS

In this paper, we investigate a core issue in the clustering of categorical data, i.e., the limitations of a single metric. Intuitively, a single metric based on certain prior knowledge cannot fully accommodate all types of datasets and clustering tasks, which unavoidably hampers clustering performance. Accordingly, we propose a novel multi-metric space fusion framework called MSF, which learns to fuse multiple metrics, enabling them to complement each other for more reasonable clustering. To better quantify the metric importance and more appropriately fuse them, normalized entropy and probabilistic metric regularization are proposed, respectively. Extensive experimental evaluation illustrates the efficacy of MSF. Despite promised merits, MSF is also not exempt from limitations. For instance, the selection of multiple metrics can be viewed as an implicit hyper-parameter to be determined by the users. A feasible solution is to simultaneously fuse a large number of metrics, which will be attempted in our further experiments. Moreover, meta-learning and continuous learning [45] could also be promising in realizing automated selection of metrics and incrementally incorporating newly developed metrics.

REFERENCES

[1] I.-E. Veliche and P. Fung, "Improving fairness and robustness in end-to-end speech recognition through unsupervised clustering," in *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[2] A. E. Ezugwu, A. M. Ikotun, O. O. Oyelade, L. Abualigah, J. O. Agushaka, C. I. Eke, and A. A. Akinyelu, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Engineering Applications of Artificial Intelligence*, vol. 110, p. 104743, 2022.

[3] Y. Zhang, Y. Zhang, Y. Lu, M. Li, X. Chen, and Y.-M. Cheung, "Asynchronous federated clustering with unknown number of clusters," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[4] D. Xu, C. Zhang, Z. Li, C. Chen, and H. Li, "Fast disentangled slim tensor learning for multi-view clustering," *arXiv preprint arXiv:2411.07685*, 2024.

[5] A. Agresti, *Categorical data analysis*, 2012, vol. 792.

[6] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[7] A. Agresti, *Analysis of Ordinal Categorical Data*. Wiley, 2010.

[8] L. Zhao, Y. Zhang, Y. Ji, A. Zeng, F. Gu, and X. Luo, "Heterogeneous drift learning: classification of mix-attribute data with concept drifts," in *Prooceedings of the 9th International Conference on Data Science and Advanced Analytics*, 2022, pp. 1–10.

[9] P. Arabie, N. D. Baier, C. F. Critchley, and M. Keynes, *Studies in Classification Data Analysis and Knowledge Organization, Cham, Switzerland: Springer*, 2006.

[10] 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*, vol. 34, no. 9, pp. 6530–6544, 2023.

[11] D. Ienco, R. G. Pensa, and R. Meo, "Context-based distance learning for categorical data clustering," in *Proceedings of the Advances in Intelligent Data Analysis VIII: 8th International Symposium on Intelligent Data Analysis, IDA 2009, Lyon, France, August 31-September 2, 2009. Proceedings 8*, 2009, pp. 83–94.

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

[13] Y.-M. Cheung and H. Zeng, "Local kernel regression score for selecting features of high-dimensional data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 12, pp. 1798–1802, 2009.

[14] Y. Zhang, Y.-M. Cheung, and K. 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.

[15] L. Bai and J. Liang, "Sparse subspace clustering with entropy-norm," in *Proceedings of the International Conference on Machine Learning*, vol. 119, 2020, pp. 561–568.

[16] D. Lin *et al.*, "An information-theoretic definition of similarity," in *Proceedings of the International Conference on Machine Learning*, vol. 98, no. 1998, 1998, pp. 296–304.

[17] J. M. Santos and M. Embrechts, "On the use of the adjusted rand index as a metric for evaluating supervised classification," in *Proceedings of the International Conference on Artificial Neural Networks*, 2009, pp. 175–184.

[18] N. Iam-On, T. Boongoen, S. Garrett, and C. Price, "A link-based approach to the cluster ensemble problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2396–2409, 2011.

[19] D. Huang, C.-D. Wang, and J.-H. Lai, "Locally weighted ensemble clustering," *IEEE Transactions on Cybernetics*, vol. 48, no. 5, pp. 1460–1473, 2017.

[20] D. Huang, C.-D. Wang, H. Peng, J. Lai, and C.-K. Kwoh, "Enhanced ensemble clustering via fast propagation of cluster-wise similarities," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 1, pp. 508–520, 2021.

[21] F. Li, J. Wang, Y. Qian, G. Liu, and K. Wang, "Fuzzy ensemble clustering based on self-coassociation and prototype propagation," *IEEE Transactions on Fuzzy Systems*, vol. 31, no. 10, pp. 3610–3623, 2023.

[22] H. Li, X. Ye, A. Imakura, and T. Sakurai, "Ensemble learning for spectral clustering," in *Proceedings of the 2020 IEEE International Conference on Data Mining*, 2020, pp. 1094–1099.

[23] J. Xu and T. Li, "Ensemble clustering with low-rank optimal laplacian matrix learning," *Applied Soft Computing*, vol. 150, p. 111095, 2024.

[24] D. Huang, C.-D. Wang, J.-S. Wu *et al.*, "Ultra-scalable spectral clustering and ensemble clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 6, pp. 1212–1226, 2020.

[25] Y. Zhang and Y.-M. Cheung, "An ordinal data clustering algorithm with automated distance learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 6869–6876.

[26] Y. Zhang, Y.-M. Cheung, and A. Zeng, "Het2hom: Representation of heterogeneous attributes into homogeneous concept spaces for categorical-and-numerical-attribute data clustering," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2022, pp. 3758–3765.

[27] L. Bai and J. Liang, "A categorical data clustering framework on graph representation," *Pattern Recognition*, vol. 128, p. 108694, 2022.

[28] 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*, vol. 44, no. 7, pp. 3560–3576, 2022.

[29] 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.

[30] Y. Zhang, M. Zhao, H. Jia, and Y.-M. Cheung, "Order is all you need for categorical data clustering," *arXiv preprint arXiv:2411.15189*, 2024.

[31] P. Wang, Y. Zhang, Y. Zhang *et al.*, "Clustering by learning the ordinal relationships of qualitative attribute values," in *Proceedings of the 2024 International Joint Conference on Neural Networks*, 2024, pp. 1–8.

[32] M. Zhao, S. Feng, Y. Zhang, M. Li, Y. Lu, and Y.-M. Cheung, "Learning order forest for qualitative-attribute data clustering," in *Proceedings of the 27th European Conference on Artificial Intelligence*, 2024.

[33] Y.-M. Cheung and H. Jia, "Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number," *Pattern Recognition*, vol. 46, no. 8, pp. 2228–2238, 2013.

[34] H. Jia and Y.-M. Cheung, "Subspace clustering of categorical and numerical data with an unknown number of clusters," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3308–3325, 2017.

[35] L. Zhao, Y. Zhang, X. Luo, Y. Zhang, Y.-M. Cheung, and K. Li, "Selecting heterogeneous features based on unified density-guided neighborhood relation for complex biomedical data analysis," in *Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine*, 2023, pp. 771–778.

[36] 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 *Proceedings of the 44th International Conference on Distributed Computing Systems*, 2024, pp. 288–299.

[37] Y. Zhang, X. Luo, Q. Chen *et al.*, "Towards unbiased minimal cluster analysis of categorical-and-numerical attribute data," in *Proceedings of the International Conference on Pattern Recognition*, 2024, pp. 254–269.

[38] Y. Zhang, R. Zou, Y. Zhang *et al.*, "Adaptive micro partition and hierarchical merging for accurate mixed data clustering," *Complex & Intelligent Systems*, vol. 11, no. 1, pp. 1–14, 2025.

[39] Z. 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.

[40] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

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

[42] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[43] A. J. Gates and Y.-Y. Ahn, "The impact of random models on clustering similarity," *Journal of Machine Learning Research*, vol. 18, no. 87, pp. 1–28, 2017.

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

[45] C. Zhang, D. Xu, X. Jia, C. Chen, and H. Li, "Continual multi-view clustering with consistent anchor guidance," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 8 2024, pp. 5434–5442.