HMVI: Clustered Natural Neighbors-Driven Heterogeneous Missing Value Inference (Supplemental Material)

TABLE I
DESCRIPTION OF THE 8 DATASETS. $d^{<n>}$, $d^{<o>}$, $d^{<u>}$ INDICATE THE NUMBERS OF NOMINAL, ORDINAL AND NUMERICAL ATTRIBUTES, RESPECTIVELY. n AND k INDICATE THE NUMBERS OF OBJECTS AND BENCHMARK CLUSTERS, RESPECTIVELY.

No.	dataset	Abbrev.	$d^{<n>}$	$d^{<o>}$	$d^{<u>}$	n	k
1	iris	IR	0	0	4	150	3
2	wine	WI	0	0	13	178	3
3	zoo	ZO	15	1	0	101	7
4	Soybean	SB	35	0	0	47	4
5	Hayes Roth	HR	2	2	0	160	3
6	Diagnosis	DS	5	0	1	120	2
7	Teacher Assistant	TA	4	0	1	151	3
8	Breast Cancer	BC	4	2	3	277	2

A. Comparison Methods

The HMVI is compared with four methods across 8 datasets, which include three types: pure numerical, pure categorical, and heterogeneous attributes. These datasets is shown in Table. I. Mean/mode substitution (MMS) is a simple method for imputing missing values in both numerical and categorical data. For numerical data, we also use K-Means clustering missing values imputation (KMCMI), which involves two steps: forming clusters with K-Means and using cluster information to handle missing values. For categorical or heterogeneous data, we use MissForest (MF), which treats the missing data problem as a prediction problem, imputing missing values by regressing each variable against all others and predicting missing values using the fitted forest. Another approach is K-Nearest Neighbors missing values imputation (KNNMI), which substitutes missing values with the mean of the k nearest complete neighbors. In the categorical experiment, simple matching serves as the core evaluation metric. As the missing rate rises, the classification accuracy of all methods generally decreases, as shown in Fig 2. Nevertheless, HMVI's classification accuracy remains higher than that of other comparative methods throughout the entire range of missing rates. Even when the missing rate is relatively high, it can still maintain relatively stable classification performance, outperforming other methods by a notable margin. This proves HMVI can better retain the key information of categorical data and ensure the reliable.

B. Evaluation and Settings

We refer to the complete data before generating missing values as the original data. We set missing rates at 10%, 20%, 30%, 40%, and 50%, and remove data completely at random for each experiment, repeating each experiment 10 times and averaging the results. To evaluate the imputation efficiency of different methods and the effect of clustering after imputation, we use two criteria. First, we assess the distance between the original and imputed data. For categorical data, we use the simple matching method for Accuracy. For numerical data, we use normalized root mean squared error (NRMSE). For heterogeneous data, we use modified RMSE(mRMSE). We evaluate the clustering effect after imputation using the Adjusted Rand Index (ARI) and the Silhouette index (CVI). We also provide clustering results for the complete original data(denoted as ORI) as a reference baseline. And in our experiment the k is set to the square root of the number of observed complete objects suggested by Lall and Sharma. Additionally, we apply KNNMI with dummy coding for categorical variables. Effective imputation methods enhance data quality and clustering performance, which in turn reflects the imputation method's quality. We will choose a classical clustering method based on the data type after imputation: K-Means for numerical data, K-Modes for categorical data, and

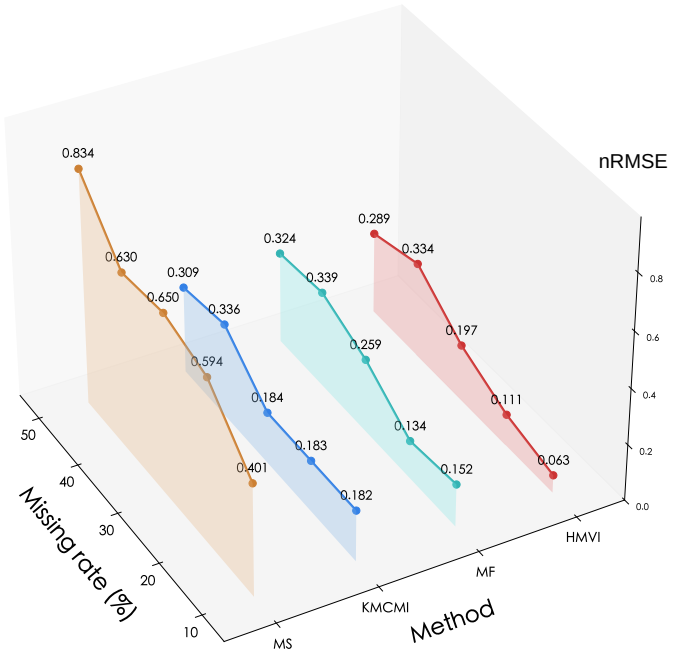

Fig. 1. Accuracy of the imputation method on pure numerical datasets.

TABLE II
CLUSTER PERFORMANCE ARI AND CVI ON NUMERICAL DATASETS.

Missing Rate	Indicator	IR					WI				
		ORI	MS	MF	KMCMI	HMVI	ORI	MS	MF	KMCMI	HMVI
10%	ARI	0.716	0.612	0.600	0.650	0.832	0.371	0.353	0.370	0.362	0.369
	CVI	0.495	0.415	0.421	0.059	0.492	0.115	0.097	0.093	0.002	0.099
20%	ARI	0.73	0.700	0.661	0.670	0.699	0.335	0.349	0.348	0.351	0.365
	CVI	0.495	0.440	0.454	0.065	0.491	0.073	0.096	0.094	0.002	0.084
30%	ARI	0.73	0.690	0.640	0.695	0.674	0.371	0.306	0.301	0.295	0.361
	CVI	0.495	0.448	0.451	0.070	0.480	0.115	0.08	0.073	0.001	0.099
40%	ARI	0.716	0.699	0.729	0.712	0.665	0.371	0.332	0.339	0.343	0.363
	CVI	0.495	0.468	0.483	0.072	0.492	0.115	0.083	0.092	0.002	0.096
50%	ARI	0.73	0.669	0.609	0.719	0.689	0.352	0.342	0.342	0.348	0.377
	CVI	0.495	0.427	0.440	0.072	0.483	0.075	0.096	0.098	0.002	0.114

TABLE III
CLUSTER PERFORMANCE ARI AND CVI ON CATEGORICAL DATASETS.

Missing Rate	Indicator	HR					ZO					SB				
		ORI	MS	MF	KNNMI	HMVI	ORI	MS	MF	KNNMI	HMVI	ORI	MS	MF	KNNMI	HMVI
10%	ARI	-0.015	0.008	0.007	-0.005	0.4652	0.688	0.675	0.652	0.606	0.831	0.824	0.596	0.741	0.651	1
	CVI	0.203	0.164	0.162	0.197	0.2586	0.493	0.368	0.345	0.366	0.498	0.397	0.248	0.360	0.288	0.482
20%	ARI	-0.002	-0.007	-0.007	-0.002	0.055	0.475	0.566	0.589	0.568	0.684	0.875	0.510	0.644	0.688	0.956
	CVI	0.219	0.233	0.201	0.221	0.223	0.382	0.291	0.258	0.284	0.475	0.428	0.217	0.321	0.301	0.496
30%	ARI	-0.013	-0.019	0.001	-0.003	0.009	0.544	0.342	0.347	0.491	0.761	0.504	0.271	0.705	0.343	1
	CVI	-0.165	0.222	0.188	0.252	0.294	0.456	0.227	0.196	0.344	0.542	0.233	0.170	0.372	0.254	0.481
40%	ARI	-0..014	0.024	0.002	-0.006	0.016	0.855	0.165	0.340	0.526	0.706	0.585	0.297	0.587	0.385	0.937
	CVI	0.216	0.290	0.163	0.255	0.274	0.491	0.189	0.195	0.360	0.655	0.271	0.180	0.324	0.247	0.526
50%	ARI	0.006	0.007	0.002	-0.004	0.011	0.602	0.121	0.238	0.203	0.637	0.591	0.057	0.616	0.214	0.758
	CVI	0.164	0.256	0.153	0.316	0.452	0.396	0.203	0.164	0.235	0.663	0.314	0.093	0.302	0.243	0.474

K-prototypes for heterogeneous data. Additionally, k is set to the true number of label classes in the dataset.

C. Imputation Accuracy Evaluation for Pure-Type Datasets

As shown in Fig 1, as the numerical datasets missing rate increases, the error metrics of different missing data handling methods all exhibit an upward trend. However, HMVI consistently maintains a relatively lower error level across all missing rate scenarios. It demonstrates that HMVI has better stability and accuracy. It can effectively reduce the interference of missing data on numerical analysis results, fully reflecting its effectiveness in handling missing numerical data.

D. Clustering Performance Validity for Pure-Type Datasets

To verify the clustering performance validity of pure-type datasets, comparative experiments on multiple typical pure-type datasets and classical numerical pure-type datasets were conducted, focusing on the correlation between imputation accuracy and clustering efficiency of different methods, and summarized the key influencing factors of clustering performance for pure-type datasets.

The clustering results of pure-type datasets are shown in Tables III and II. A key advantage of HMVI is that it integrates imputation into the clustering process. This end-to-end design avoids the error accumulation caused by "separate imputation then clustering" in traditional methods, thus achieving excellent clustering efficiency on pure-type datasets.

In general, for pure-type datasets, better imputation accuracy leads to better clustering results. This is because pure-type datasets have a relatively unified data distribution, and accurate imputation can retain the original distribution characteristics, providing a reliable basis for clustering. However, an exception was observed in the HR pure-type dataset: the MS method had low imputation accuracy, but its clustering efficiency was unexpectedly high. Further analysis shows that for pure-type datasets with a large sample size (such as the HR dataset), when a large amount of data is incorrectly imputed, the erroneous data will form relatively concentrated "pseudo-clusters". This is due to the single original distribution of pure-type datasets, which makes wrong values more likely to show aggregation characteristics. This aggregation of erroneous data mistakenly improves the clustering evaluation

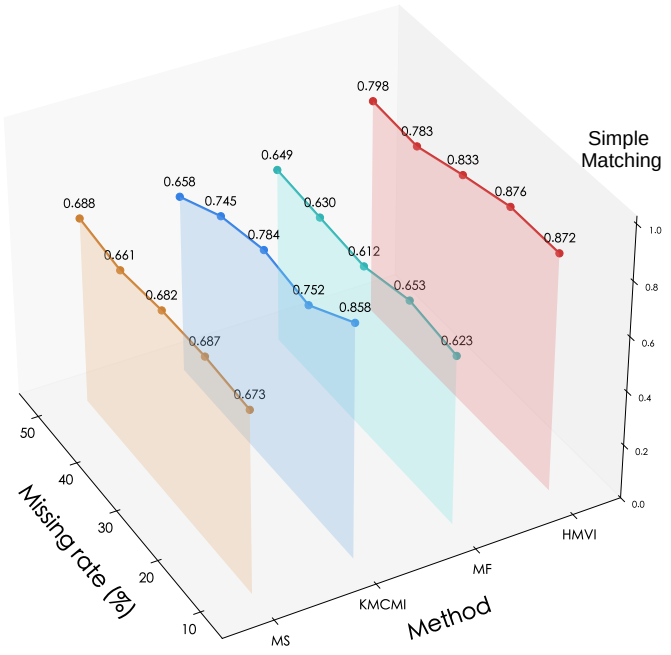

Fig. 2. Accuracy of the imputation method on pure categorical datasets.

indicators, which is a special phenomenon worthy of attention in the clustering of pure-type datasets.

For other methods (such as MF and KNNMI) on pure-type datasets, a consistent rule was also found: as the imputation error decreases, their clustering effects gradually improve. This further proves that correctly imputing pure-type datasets has a significant impact on clustering results. Since the data distribution of pure-type datasets is relatively simple, even small imputation errors may distort the distribution, while reducing errors can better restore the true data structure and promote effective clustering.