Order Is All You Need for Categorical Data Clustering

Author Response to Reviewer SKH2
Applying OCL to Numerical Data

Anonymous Author(s)

Experimental Settings. To demonstrate the potential of our OCL for numerical data clustering, we compare OCL on two numerical datasets obtained from the UCI machine learning repository (http://archive.ics.uci.edu/ml). The dataset statistics are provided in Table 1. To implement OCL on these two numerical datasets, we adopt uniform discretization for the numerical attributes. That is, the values of each numerical attribute are divided into ten intervals of equal length, and these ten intervals are regarded as ten possible values with an order relationship. We implement OCL on the discretized numerical datasets to learn new orders for the intervals. To illustrate the effectiveness of order learning, we also preserve the original order by implementing OCL without order learning (OCL w.o. OL). More specifically, OCL w.o. OL is a version of OCL with fixed original orders. It is equivalent to executing only the inner loop of Algorithm 1 in our paper to learn clusters. The outer loop does not make any adjustments to the order to ensure that it is fixed. The conventional K-Means clustering algorithm that utilizes the original order of numerical data by default is also compared. All the results are averaged on ten runs of the compared methods.

Clustering Performance. The comparative results are shown in Table 2. It can be seen that the order learned by OCL can surpass the original order preserved by the K-Means and OCL w.o. OL in general. This is because the order of a numerical attribute is not necessarily linearly related to the true clusters. For example, people with medium incomes may have happier family lives than those with very high or meager incomes. Therefore, the original order of the "Income" attribute may not be beneficial for forming two clusters of people with happy and unhappy family lives. In this scenario, the proper income order should be something like: medium > high > low, or reverse. The order learned by OCL can reflect such non-linear relationships among values, and thus yields more accurate clustering results in the corresponding situations.

Detailed Observations. Three detailed observations are provided in the following: (1) OCL performs better than the two counterparts with fixed original orders (i.e., OCL w.o. OL and K-Means) in general. This proves the effectiveness of the orders learned by OCL on the two numerical datasets. This also verifies our guess that order learning can be extended to numerical data clustering. By carefully observing the data attribute values and cluster distributions, we found that the nonlinear relationship between attributes and the label attribute is the key factor causing this phenomenon. However, the main challenge lies in that the labels of the data for clustering are all categorical. How to reflect the implicit relationship between heterogeneous numerical attribute and categorical label attribute is the key to deploying order learning for numerical data clustering. (2) Both K-Means and OCL w.o. OL use the original order of numerical attributes, but the latter performs better in general. This is because the latter OCL w.o. OL uses a more advanced learnable distance, which adaptively adjusts the distances between attribute values during clustering. In contrast, K-Means uses the Euclidean distance, which is given in advance and thus has no connection with the clustering task. (3) The NMI performance of K-Means is obviously higher than that of the two OCL variants (i.e., the full version of OCL and OCL w.o. OL) on the Shuttle dataset. By observing the distributions of attribute values and clusters in detail, we found that the distribution of the shuttle dataset is extremely imbalanced. The values of many attributes are concentrated in a very small interval, which is very unfavorable for the equal interval discretization adopted by the two OCL variants. Therefore, in comparison, the Euclidean distance used by K-Means can retain more discrimination information between samples, so the corresponding clustering performance is better in terms of the NMI metric. On the other hand, the cluster distribution of this dataset is also extremely imbalanced (the imbalance ratio, i.e. the number of samples in the largest true cluster divided by the number of samples in the smallest true cluster, is as high as 5684.7!). The extreme imbalance will lead to unstable verification results of the ARI and NMI metrics. This explains why the compared methods exhibit very different performance positions on ARI and NMI. However, since the CA metric first performs the optimal mapping from the obtained clusters to the label clusters and then evaluates the performance, CA will be relatively reliable in the extremely imbalanced case and thus still verifies the effectiveness of order learning of numerical data.

Table 1: Dataset statistics. d, n, and k^* are the numbers of attributes, samples, and true clusters, respectively.

No.	Dataset	d	n	k^*
1	Shuttle	9	43500	7
2	Avila	10	20867	12

Table 2: Clustering performance on two numerical datasets. The best results on each dataset are highlighted in bold.

Data	Metric	K-Means	OCL w.o. OL	OCL
Shuttle	CA	0.4217±0.00	0.5358±0.03	**0.6415±0.02**
	ARI	0.1912±0.00	0.2234±0.02	**0.2640±0.01**
	NMI	**0.4349±0.00**	0.3566±0.01	0.3115±0.01
Avail	CA	0.2239±0.00	0.2487±0.00	**0.3212±0.00**
	ARI	0.0201±0.00	0.0301±0.00	**0.0323±0.00**
	NMI	0.0854±0.00	0.0920±0.00	**0.0953±0.00**