Towards Unbiased Minimal Cluster Analysis of Categorical-and-Numerical Attribute Data

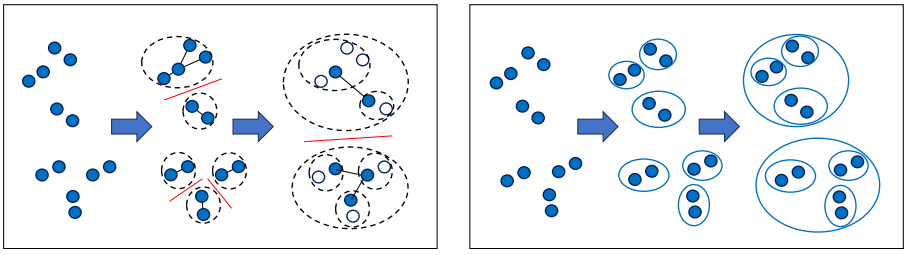
— Supplementary Materials —

Yunfan Zhang[1], Xiaopeng Luo[1], Qingsheng Chen[1], Rong Zou[2], Yiqun Zhang[1⋆], and Yiu-ming Cheung[2]

[1] Guangdong University of Technology, Guangdong, China
[2] Hong Kong Baptist University, HongKong, China
3121008002@mail2.gdut.edu.cn, gordonlok@foxmail.com, 2112205080@mail2.gdut.edu.cn, rongzou@comp.hkbu.edu.hk, yqzhang@gdut.edu.cn, ymc@comp.hkbu.edu.hk

1 Comparison of Mic2Mac and Hierarchical Clustering

As hierarchical clustering is similar to ours, the differences between our Mic2Mac and hierarchical clustering are demonstrated in Fig. 1 (a) and (b), respectively. In Fig. 1 (a), the regions are delineated by black dashed circles, and partitioned by red solid lines. Representative objects in each region are located at the points where the black lines intersect, and merging is constrained by the merging interval. In Fig. 1 (b), objects or clusters undergo merging at each layer. The typical approach involves pairwise combining one by one, contrasting with hierarchical merging where multiple clusters can be merged simultaneously.



(a) (b)

Fig. 1: Comparison of clustering process between Mic2Mac algorithm (a) and hierarchical clustering (b).

⋆ Corresponding author

2 Proof of Theorem 2

Theorem 2. *The time complexity of* Mic2Mac is $O(d^2n + n\sqrt{\log n})$ with any time-step τ.

Proof. To analyze the worst case, we assume all attributes are categorical, i.e., $d = |A|$, and V is equal to the maximum number of possible values across all the categorical attributes. Given the distance matrices $D^{(\tau)}$ in Algorithm 1 that correspond to the attribute in A beforehand, we need to partition the micro-clusters $M^{\phi,(\tau)}$. Moreover, the time complexity is equal in every time-step. To analyze the time complexity, we compute first $D^{(\tau)}$ and $M^{\phi,(\tau)}$ once, respectively.

To compute $D^{(\tau)}$, we need to derive $d \times d$ pairs of CPDs by scanning n data objects in data set S one time. This results in a $O(nd^2)$ complexity. For computing the distances between a pair of intra-attribute possible values in Eq. (7), it takes $O(V)$ complexity for every attribute. To determine the weights based on Eq. (8), it needs $V(V-1)/2$ calculations for a total time complexity of $O(V^2)$ for every attribute. To establish the weights between each pair of d attributes, we are required to compute $V(V-1)/2$ distances for each attribute. Thus, obtaining $D^{(\tau)}$ incurs a complexity of $O(nd^2 + V^3d^2 + V^2d^2)$, which can be simplified to $O(nd^2 + V^3d^2)$.

Given the $D^{(\tau)}$ obtaining from Algorithm 1,to compute $M^{\phi,(\tau)}$, we need to concurrently sort an $n \times n$ matrix, taking $O(n + n\sqrt{\log n})$ complexity. Then, for each of the n data objects, we conclude the q_i before calculating its density ρ_i using Eq. (11) by concurrently comparing it to the remaining $n-1$ objects, with a $O(n)$ time. Moreover, we concurrently search for remaining at most $n-1$ objects for each object to acquire the one with greater density and the closest distance to it to obtain n merging intervals, which takes $O(n)$ time. Subsequently, we sort n merging intervals in $O(n\sqrt{\log n})$ complexity. Form n neighborhood sets in order, each considering at most q_i objects from the already arranged distances, taking $O(nq_i)$ complexity. The overall complexity for updating $M^{\phi,(\tau)}$ is $O(3n + 2n\sqrt{\log n})$.

Given the micro representative objects set MR, we need to update S in each round, which takes n in the worst-case. Therefore, the time complexity for hierarchical merging is $O(n)$ in each round.

Therefore, the overall complexity of Mic2Mac at a given time-step τ can be expressed as $O(nd^2 + V^3d^2 + n\log n + n)$. Since V is usually a small invariable, typically ranging [1, 5], which means that $V^3 \ll n$ for most real-world data sets, the ultimate complexity of Mic2Mac can be simplified to $O(d^2n + n\sqrt{\log n})$. □

3 Proof of Theorem 3

Theorem 3. *Eq.(8) provides a consistent treatment of both categorical and numerical attributes, assuming that each numerical attribute is an independent one-dimensional continuous Euclidean distance space with a domain of [0,1].*

Proof. Each numerical attribute can be considered as linear distance space with an infinite number of evenly spaced possible values. For numerical attributes, w^{rt} is always equal to 0 when $r \neq t$, as they are independent. Therefore, we only need to consider the case when $r = t$, where $w^{rt} = \lim_{j \to \infty} (\sum_{i=1}^{j} 1/j)/1 = 1$ according to Eq. (9). The $j \to \infty$ represents the number of intervals between adjacent possible values, and the denominator "1" denotes only one inter-concept distance for the attribute, since there are only two concepts, "0" and "1". Thus, when $r = t$, the EMD with Euclidean distance is downgraded to $D^{rt}(v_h^r, v_o^r) = |v_h^r - v_o^r|$, which equals to the Euclidean distance in $D(\mathbf{x}_i, \mathbf{x}_j)$ as described in Eq. (5). □