Order Is All You Need for Categorical Data Clustering

Author Response to Reviewer BSHc
Problem Formulation and List of Symbols

3.1 Problem Formulation

The problem of categorical data clustering with order learning is formulated as follows. Given a categorical dataset $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ with n data samples (e.g., a collection of n clients of a bank). Each data sample $\mathbf{x}_i$ can be denoted as a d-dimensional row vector $\mathbf{x}_i = [x_{i,1}, x_{i,2}, .., x_{i,d}]^\top$ represented by d attributes $A = \{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_d\}$ (e.g., occupation, gender, priority, etc.). Each attribute $\mathbf{a}_r$ (e.g., occupation) can be denoted as a column vector $\mathbf{a}_r = [x_{1,r}, x_{2,r}, ..., x_{n,r}]$ composed of the occupation description of all the n clients. For the categorical attribute "occupation" $\mathbf{a}_r$, all its n values are taken from a limited number of qualitative possible values $V_r = \{v_{r,1}, v_{r,2}, ..., v_{r,l_r}\}$, e.g., {lawyer, doctor, ..., scientist}. The subscripts indicate the sequential number of an attribute and the sequential number of the attribute's possible value. For example, doctor is the second value in the r-th attribute "occupation", so we have $v_{r,2}$ = doctor. We use l_r to indicate the total number of possible values of an attribute, and it usually satisfies $l_r \ll n$ for real categorical data. Please note that the sequential numbers of possible values only distinguish them, and do not indicate their order.

Assume that there is an optimal order indicating an appropriate underlying distance structure of possible values w.r.t. a clustering task. We denote the optimal order of an attribute $\mathbf{a}_r$ as $O_r^* = \{o^*(v_{r,1}), o^*(v_{r,2}), ..., o^*(v_{r,l_r})\}$ where the superscript "$*$" marks the optimum and the value of $o^*(v_{r,g})$ is an integer reflecting the unique ranking of $v_{r,g}$ among all the l_r possible values in V_r. By still taking the "occupation" attribute as an example, if its three possible values $v_{r,1}$ = lawyer, $v_{r,2}$ = doctor, and $v_{r,3}$ = scientist, ranking second (i.e., o(lawyer) = 2), third (i.e., o(doctor) = 3), and first (i.e., o(scientist) = 1), respectively. Then the corresponding order of these values are $O_r = \{o(\text{lawyer}), o(\text{doctor}), o(\text{scientist})\} = \{2, 3, 1\}$. Such order reflects a distance structure that roughly satisfies $\text{dist}(v_{r,g}, v_{r,h}) \propto |o(v_{r,g}) - o(v_{r,h})|$. More specifically, the following distance relationship should hold:

(1) dist(doctor, scientist) > dist(doctor, lawyer),
(2) dist(doctor, scientist) > dist(lawyer, scientist),

because "doctor" and "scientist" are further away in order, compared to "lawyer" and "doctor" that are adjacent in order.

We aim to approximate the optimal orders of all the d attributes $O^* = \{O_1^*, O_2^*, ..., O_d^*\}$. Then the order information reflecting the distance structure can guide the partition of data objects during clustering. Our approach, i.e., learning order relations, is more generative than most existing approaches that directly define the distance between attribute values. We focus more on how the intrinsic relationship between values affects the distance between samples in forming clusters, rather than simply defining the clustering-irrelevant distance between values based on data statistics, e.g., occurrence probability of values in an attribute.

Table 1: Frequently used symbols. Note that we uniformly use lowercase, uppercase, bold lowercase, and bold uppercase to indicate value, set, vector, and matrix, respectively.

Symbol	Explanation
X, A, O, and C	Dataset, attribute set, order set, and cluster set
$\mathbf{x}_i$ and $x_{i,r}$	i-th data object and r-th value of $\mathbf{x}_i$
$\mathbf{a}_r$, V_r, and $v_{r,g}$	r-th attribute, its value set, and g-th value of V_r
l_r	Total number of possible values in V_r
O_r	Order values (integers) of $\mathbf{a}_r$'s possible values V_r
$o(v_{r,g})$	Order value (integer) of possible value $v_{r,g}$
$\mathbf{Q}$ and C_m	Object-cluster affiliation matrix and m-th cluster
$q_{i,m}$	(i,m)-th entry of $\mathbf{Q}$ indicating affiliation of $\mathbf{x}_i$ to C_m
$\Theta(\mathbf{x}_i, C_m; O)$	Overall order distance between $\mathbf{x}_i$ and C_m
$\theta(x_{i,r}, \mathbf{p}_{m,r})$	Order distance between $\mathbf{x}_i$ and C_m reflected by $\mathbf{a}_r$
$\mathbf{D}$ and $\mathbf{d}_{i,r}$	Order distance matrix and its (i,r)-th entry
$d_{i,r,g}$	Order distance between $x_{i,r}$ and $v_{r,g}$, also the g-th value of $\mathbf{d}_{i,r}$
$\mathbf{P}$ and $\mathbf{p}_{m,r}$	Conditional probability matrix and its (m,r)-th entry
$p_{m,r,g}$	Occurrence probability of $v_{r,g}$ in C_m, also the g-th value of $\mathbf{p}_{m,r}$
$O_{r,m}$	Order values of $\mathbf{a}_r$'s possible values V_r obtained through C_m
$L_{r,m}$	A fraction of L jointly contributed by $\mathbf{a}_r$ and C_m

For a crisp partitional clustering task to partition X into k non-overlapping subsets $C = \{C_1, C_2, ..., C_k\}$, the objective can be formalized as the problem of minimizing:

$$L(\mathbf{Q}, O) = \sum_{m=1}^{k} \sum_{i=1}^{n} q_{i,m} \cdot \Theta(\mathbf{x}_i, C_m; O) \tag{1}$$

where $\mathbf{Q}$ is an $n \times k$ matrix with its (i,m)-th entry $q_{i,m}$ indicating the affiliation between sample $\mathbf{x}_i$ and cluster C_m. Specifically, $q_{i,m} = 1$ indicates that $\mathbf{x}_i$ belongs to C_m, while $q_{i,m} = 0$ indicates that $\mathbf{x}_i$ belongs to a cluster other than C_m, which can be written as:

$$q_{i,m} = \begin{cases} 1, & \text{if } m = \arg\min_y \Theta(\mathbf{x}_i, C_y; O) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

with $\sum_{i=m}^{k} q_{i,m} = 1$. The distance $\Theta(\mathbf{x}_i, C_m; O)$ reflects the dissimilarity between $\mathbf{x}_i$ and C_m computed based on the distance structures reflected by the orders O. Therefore, the problem to be solved can be specified as how to jointly learn $\mathbf{Q}$ and O to minimize L. When the algorithm converges, a sub-optimal solution with k compact bank client clusters is expected to be obtained. In general, "compact" means that the bank clients with more similar characteristics (i.e., with more similar value descriptions across all the attributes) are gathered into the same cluster, and different clusters are relatively distinguishable.

Frequently used symbols in this paper and the corresponding explanations are sorted out in Table 1 below. Then we discuss how to define the order distance and learn order for accurate categorical data clustering.