Dissimilarity Clustering by Hierarchical Multi-Level Refinement

Brieuc Conan-Guez, Fabrice Rossi
brieuc.conan-guez@univ-lorraine.fr, Fabrice.Rossi@univ-paris1.fr

Dissimilarity Clustering

We aim at clustering the object set \( \Omega \) (with \( |\Omega| = N \)) into \( c \) clusters described by the partition \( P = \{ C_1, \ldots, C_c \} \) using only a reflexive and symmetric dissimilarity measure \( d \) on \( \Omega \). To achieve this goal, we try to minimize the following generalization of the standard K-means quantization error to this setting, that is [1]:

\[
E(P) = \sum_{C_i \in P} \frac{1}{|C_i|} \sum_{i \in C_i} d(i,j)
\]

This can be done greedily using the standard hierarchical clustering analysis (HCA) technique, based on the following linkage criterion:

\[
\Delta E(C_p, C_q) = \frac{S_{pq}}{|C_p| + |C_q|} - \frac{S_p}{|C_p|} - \frac{S_q}{|C_q|}
\]

where \( S_{pq} = \sum_{i \in C_p, j \in C_q} d(i,j) \)

Fast Hierarchical Clustering

HCA is implemented using the efficient algorithm proposed by D. Müllner for agglomerative hierarchical clustering [2]. Even though this recent algorithm has a worst case complexity of \( O(N^3) \), it is much more efficient than actual commonly used algorithms in practical situations where its observed scaling is in \( O(N^2) \). Moreover this algorithm can handle general linkage criteria, such as the one defined above \( \Delta E \).

Partition Refinement

A drawback of greedy methods is that bad merging decisions done during the early steps of the HCA cannot be corrected in the later steps, which leads to a suboptimal partition. To limit this problem, a classical approach consists in refining the partition (the chosen level of the hierarchy) thanks to a greedy heuristic. The refinement performs a local search by moving one object at a time from one cluster to another.

In the context of hierarchical clustering, such heuristics are known as Single-Level Refinement (SLR) approaches, as they operate on just one level of the hierarchy: the bottom level.

Multi-Level Refinement

Contrarily to SLR approaches, the Multi-Level Refinement (MLR) approach [3, 4] operates on different levels of the hierarchy. Let \( P_K \) be the level chosen by the analyst (the partition to refine). Given a reduction factor \( \alpha < 1 \), the MLR considers only a small number of levels \( \{P_0, \ldots, P_K\} \) of the hierarchy. The selected levels are those whose sizes are given by a geometric progression \( (N, \alpha N, \alpha^2 N, \ldots, \alpha^{K-1} N, |P_K|) \). This series begins by the partition of singletons \( P_0 (size N) \), and ends by the partition to refine \( P_K \) (size \( |P_K| \)).

The MLR proceeds in a top-down way. It first applies the greedy refinement heuristic to \( P_K \) considered as a partition of clusters of partition \( P_{K-1} \): this first refinement corresponds to moving entire sub-clusters of objects from one cluster of \( P_K \) to another one. Once this is done, the modifications of \( P_K \) are projected onto \( P_{K-2} \). As a second step, it refines \( P_K \) considered as a partition of clusters of \( P_{K-2} \) and projects once again the modifications. The process repeats identically for all the levels below (i.e. for \( k \) varying from \( K - 3 \) to 0). The final step \( (k = 0) \) corresponds to applying the greedy refinement on the objects as in the SLR.

Experiments

The proposed method is tested on two classical dissimilarity data sets: the small size cat cortex database with 65 objects and the large size Copenhagen chromosome database with 4200 objects. Reference performances are provided by the relational k-means (RKM). In all cases, the RKM is started from a number of initial random configurations chosen so that both methods use approximately the same computational resources.

Quantization error \( E \) as a function of the number of clusters for two databases (cat cortex on the left and chromosome on the right) - top solid line: standard HCA, bottom solid line: MLR, dashed lines: best and worst results of the RKM.

References