My Data Science Challenging Problems # 52: Clustering and Labeling with Minimal Human Computer Interactions

*** *** ***,***

XXX

XXXX

xxx

November 16, 2018

Abstract

If you happen to have a solution of this question, please contact Haoda Fu at fu_haoda@lilly.com or fuhaoda@gmail.com or commit your solution to `https://github.com/fuhaoda/DCSP52_HumanComputerInteractionClustering.git`

Key words and Phrases: *** ***.
Short title: ***

1 Introduction

Clustering analysis (such as k-mean, hierarchical clustering) algorithms are useful to understand the association among subjects. To apply those algorithms, one fundamental question is to define the similarity and dissimilarity between different subjects, i.e. to define $\|X_i - X_j\|$.

For a fixed population, $\{X_1, \cdots, X_n\}$, using default clustering algorithms, computer may generate one way of clustering the subjects, and the results may not be the desired results as human expected. We would like to have smallest number of human and computer interactions, so that we can let computer to figure out what are the human desired clustering results. Details are listed in the next section.

2 Problem details

Suppose we have subjects $\{X_1, \cdots, X_n\}$. For each subject X_i, we have multiple attributes, i.e. $X_i = (X_{i1}, \cdots, X_{ip})$ which is a p dimensional vector. Let $d(X_i, X_j) = \{\sum_{k=1}^{p} w_k(X_{ik} - X_{jk})^2\}^{1/2}$ be the human defined norm which has unknown parameters $\{w_1, \cdots, w_k\}$. We would like to estimate $\{w_1, \cdots, w_k\}$ through human and computer interaction from the following steps.

1. Start from a guess, say $w_1 = w_2 = \cdots = w_k = 1$, and apply clustering method, say k-mean algorithm, to generate one clustering result A.

2. Generate another set of weights, and apply the same clustering algorithm (using new weights). Then, we have another clustering result B.

3. Ask human to compare result A vs result B to choose a better one, and keep the better result to call it result A, then go to step 2.

4. Iterate step 2 and step 3 until that the human is happy with the result A.

How can we generate weights in step 2, so that we can have smallest number of iterations?

This algorithms may be related to convex optimization.

3 Potential Applications

We often have large amount of unlabeled data. It is infeasible to let human to label each of them. So we need an algorithm to get majority of the label correct with fewest human and computer interactions. Then we can let human expert to figure out the labels in the classification boundary.

References