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

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

XXX

XXXX

xxx

April 16, 2023

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

The field of machine learning is faced with a pressing challenge - an enormous amount of unlabeled data, particularly in the medical field. According to estimates, less than 1% of medical images are currently labeled and ready for supervised machine learning tasks. Labeling data can be time-consuming and requires domain knowledge from experts who may have limited availability. Therefore, it is crucial to develop algorithms that can efficiently use human expert time to label the data. To address this challenge, investment in research and development is essential. Effective algorithms can leverage the expertise of human annotators more efficiently and unlock the vast potential of the available unlabeled data.

Recent research has shown promising results in this area, particularly in the use of active learning, semi-supervised learning, and transfer learning (e.g. our recent JASA paper click here).

We are currently seeking an algorithm capable of handling high-dimensional data such as images, voices, texts, and videos. To make this problem more concrete, we can use the MNIST database as a testbed (click here for reference). This large labeled database contains 60,000 training images and 10,000 testing images of handwritten digits. Let's remove all the labels for those 60,000 training data to fit our purpose. Suppose we have only 10 minutes for a person to label the 60,000 images accurately. A naive solution would be to label images one by one, which would take approximately 10 minutes per 600 images. Then, we could train a deep learning algorithm using these 600 labeled images and predict the labels for the remaining 59,400 images. This approach could result in around 80% accuracy on the entire dataset. How can we do better?

We may quickly indentify some limitations of this naive approach:

- This naive approach randomly selected those 600 images and query the human one by one. Can we enable the computer to ask the most relevant questions based on what the algorithm has already learned?
- Labeling one image at a time is not the only query method available. We could also ask whether two images belong to the same class or what the majority class is among 11 images. Each query method (Q_i) is associated with a cost (C_i) and potentially the probability of correct answers (p_i). Therefore, we have a choice of query as $\mathcal{Q} = \{(Q_i, C_i, p_i), i = 1, \cdots, k\}$.

2 Problem details

One potential solution that we are exploring is learning a metric that aligns with our labeling objective.

In the ideal case, we could directly apply a non-supervised clustering algorithm, such as the K-means method, and obtain 10 classes for the handwritten zip codes that perfectly correspond to the digits from 0 to 9. Unfortunately, this is not typically the case. The reason is that clustering analysis algorithms, such as K-means and hierarchical clustering, rely on the similarity and dissimilarity between different subjects, i.e. a metric. However, the metric used in clustering methods may not align with our objective of labeling. Therefore, we are investigating methods to iteratively learn a metric that better aligns with our objective and can improve the accuracy of labeling.

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 $d(X_i, X_j) =$

$\{(X_i - X_j)^\top A(X_i - X_j)\}^{1/2}$ be the human defined norm which has unknown parameters in a semi-positive definite matrix A. We would like to estimate A through human and computer interaction from the following steps. The collection of $\mathcal{A} = A$ can be considered as a space.

1. Start from a guess, say $A = I$, or some A comes from transfer learning.
2. Choose a query method from $\mathcal{Q}$ to query a human.
3. The result can be denoted as $f(A)$ to measure how good of the current A.
4. Iterate step 2 and step 3 until that the human is happy with the result A.

We need to consider a few things in the above approach, for example,

- How can we know how complex of the space and what is the current status?
- How can we develop a method to choose different types of query? and generate most informative samples corresponding to this query?
- How can we develop an optimization algorithm on space $\mathcal{A}$?