Online Active Metric Learning for Clustering

Parag Jain, Vineeth N Balasubramanian
cs13m1008@iith.ac.in, vineethnb@iith.ac.in

Introduction

1. Most clustering algorithms rely on the definition of a suitable distance metric.
2. Few methods have been proposed to update a learned metric when newer data is available. But with exponential growth of data, it is difficult to update a learned metric on continual basis with all new data.
3. This creates a need for actively selecting data points from newer data that can be used to update the learned metric.
4. We propose an online distance metric learning method for the clustering setting, that can actively select informative pairs of points from the given data to update the metric.

Proposed: Online active metric learning

Select Pairs: Select pairs algorithm is a heuristic based approach to select pair of points using current clustering setup. Selected pair of points are posed as a query to the user to label them as similar or dissimilar.

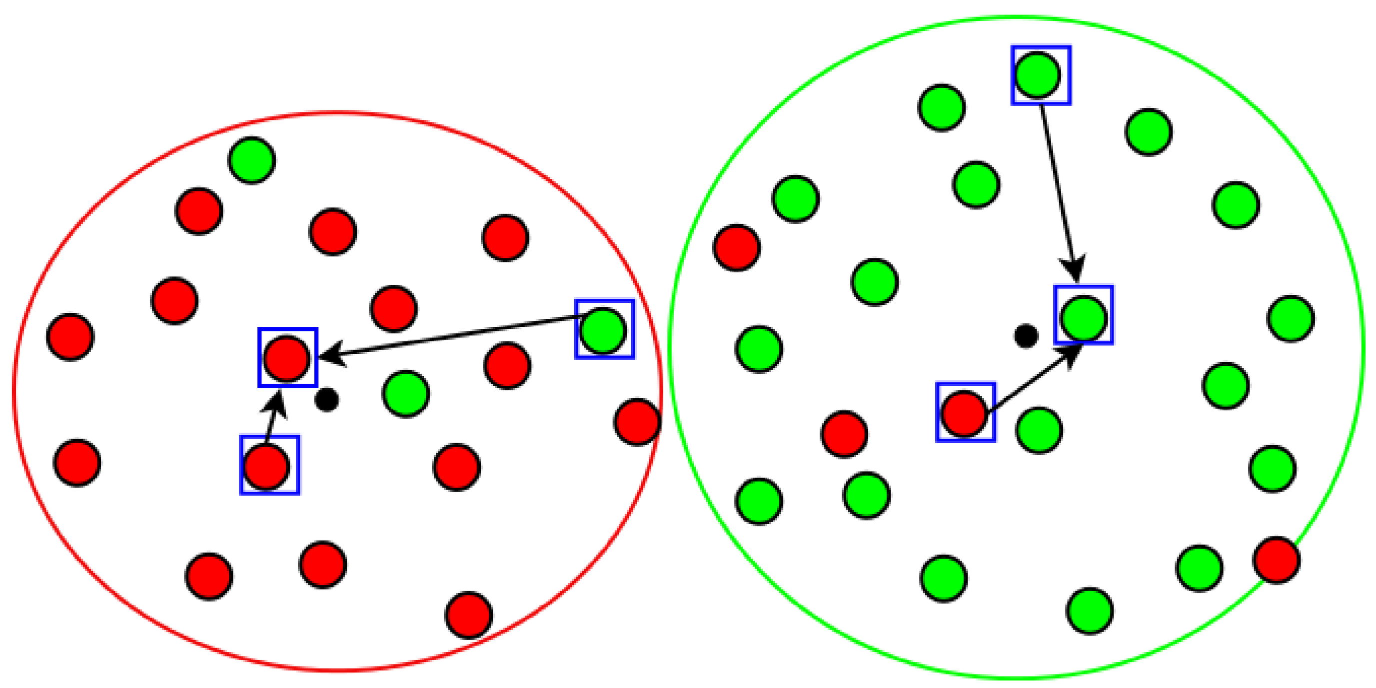


Figure 1: Schematic illustration of actively selecting pairs

Algorithm 1 Online active metric learning

procedure OAML
  Input: X, p, r, α
  k number of cluster
  N number of iterations
  $X_{n\times d}$ is input feature matrix
  r is ratio of near-center & boundary points to select
  p is number of pairs to select in each iteration
  α is alpha-trimming parameter
  $[S, D]$ similar & dissimilar pairs
  Output: A: Learned metric ; $C_{n\times k}$: Cluster belongingness matrix
  Initialization: $A \leftarrow I$
  for i = 1 to N **do**
    $\hat{C} \leftarrow kmeans(X, k, A)$
    $[S, D] \leftarrow SELECTPAIRS(X, C, p, S, D, A, r, \alpha)$
    Update matrix A based on new pairs using POLA, $A \leftarrow pola(A, S, D)$
  end for
end procedure

Pseudo-Metric Online Learning

- Learn metric of form: $d_A(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_j - \mathbf{x}_i)^T A(\mathbf{x}_j - \mathbf{x}_i)}$
- Given initial matrix threshold pair (A_τ, b_τ).
- On receiving a pair of instance and corresponding label $z = (x, x', y)$, POML updates matrix threshold pair to $(A_{\tau+1}, b_{\tau+1})$. Such that, $A_\tau \succeq 0$ and $b_\tau \geq 1$
- Prediction loss is defined as,

$$l_\tau(A, b) = max\{0, y_\tau(d_A(x, x')^2 - b) + 1\}$$

- Update matrix-threshold pair by projecting into C_L & C_A where,

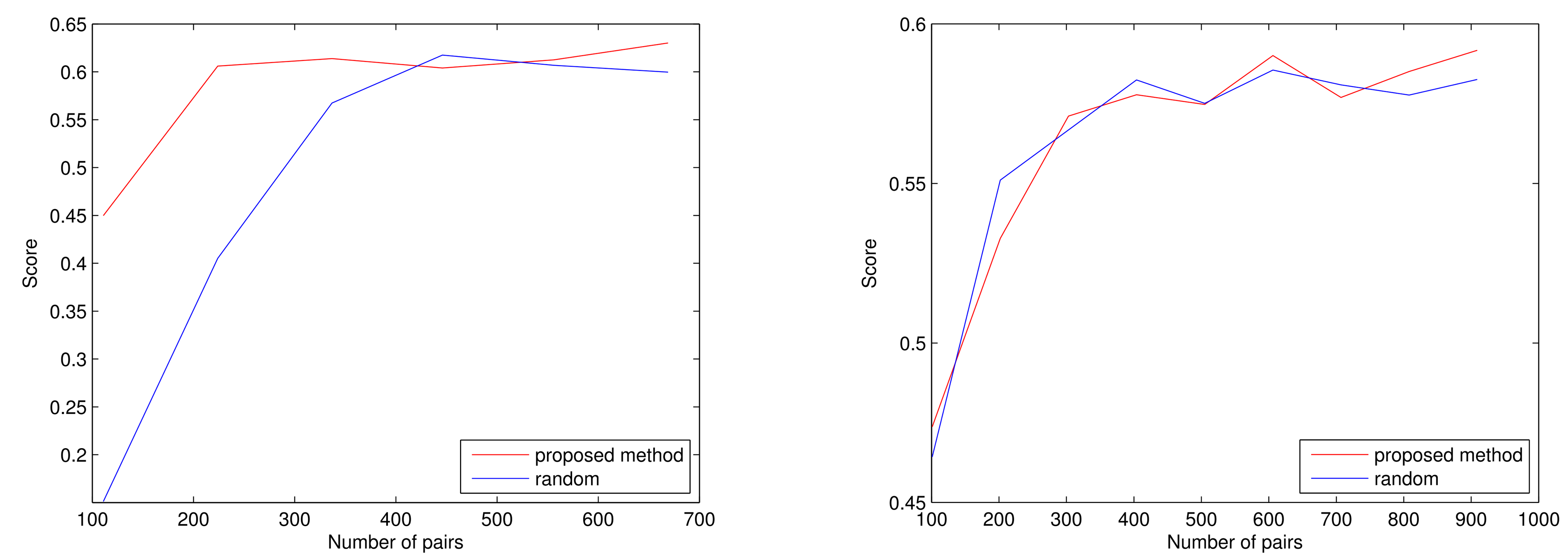
$$C_\tau = \{(A, b) \in \mathbb{R}^{(n^2+1)} : l_\tau(A, b) = 0\} \quad C_a = \{(A, b) \in \mathbb{R}^{(n^2+1)} : A \succeq 0, b \geq 1\}$$

Why pairwise clustering?

- Easier for a user to determine whether two points are similar or not, rather than providing true labels.
- Pairwise constraints relates better to the clustering problem.

Results

We used the Silhouette co-efficient measure to validate the clustering. This is given as $S = \frac{\sum_i s(i)}{n}$, where, $s(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}}$. Label complexity (the number of labellings required to reach a predefined score) for our method is significantly lesser, which is what we aim here.



(a) Silhouette score for Letter recognition dataset

(b) Silhouette score for Magic dataset

Future work

Our future work is to come up with a single-stage integrated approach for online active metric learning, instead of the proposed two-stage approach.