1 Introduction

We consider the problem of learning a measure of distance among feature vectors, and propose a hybrid method that simultaneously learns from similarity ratings and class labels. Application: information retrieval.

2 Problem Formulation

- Pair-wise similarity rating: a set of quintuplets \((o, o', x, x', o)\)
  - \(o, o':\) object identifier
  - \(x, x' \in \mathbb{R}^k: \) features of each object
  - \(\sigma \in \{1, 2, 3\}: \) dissimilar / neutral / similar
- Class labels: a set \(O\) of triplets \((o, x, c)\)
  - \(c \in \{1, 2, \ldots, M\}: \) class
- Distance Metric:
  \[
  d_u(x, x') = \sum_{i=1}^{k} r_i (x_i - x_i')^2
  \]
- Main Question: how to learn coefficients \(r\)?

3 Conventional Algorithms

**Ordinal Regression:** We assume
\[
P(\sigma \leq |x, x'|) = \frac{1}{1 + \exp(-d_u(x, x') - \theta_i)}
\]
where \(i\) is the level of similarity, and solve
\[
\max_{r \geq 0, \theta_1 \leq \theta_2} \sum_{i=1}^{k} \log P(\sigma|x, x')
\]
subject to \(r \geq 0\) and \(\theta_1 \leq \theta_2\).

**Convex Optimization:** We solve
\[
\min_{r \geq 0} \sum_{i=1}^{k} d_u^2(x, x')
\]
subject to \(d_u(x, x') \geq 1\) and \(r \geq 0\).

**Neighborhood Component Analysis:** We assume a feature \(x^t\) is assigned class label \(c^t\) with probability
\[
P(c^t|x^t, G) = \frac{\exp(-d_u(x^t, x^t))}{\sum_{x^t'} \exp(-d_u(x^t', x^t'))}
\]
and solve
\[
\max \sum_{i=1}^{k} \log P(c|x, G(x, c))
\]

4 A Hybrid Method

**Assumptions:**
- The observed features may not express all relevant information. In other words, there are "missing features."
- Given objects \(o, o'\) with observed feature vectors \(x, x' \in \mathbb{R}^k\) and missing feature vectors \(z, z' \in \mathbb{R}^l\), the underlying distance metric is given by
  \[
  D(o, o') = \left( \sum_{i=1}^{k} r_i (x_i - x_i')^2 + \sum_{j=1}^{l} r'_j (z_j - z'_j)^2 \right)^{1/2}
  \]
where \(r \in \mathbb{R}^k\) and \(r' \in \mathbb{R}^l\).
- \(x\) and \(z\) are independent conditioned on \(c\).

**Algorithm:**

We can use the above results into a learning algorithm that learns the coefficients of a distance metric from feature differences and similarity ratings.

Example: We take \(o\) to be a kernel density estimator similar to NCA, and \(B\) to be the aforementioned convex optimization:
\[
\min_{r \geq 0} \sum_{i=1}^{k} d_u^2(x, x') + \lambda_u^2 Qu(o')
\]
subject to \(r \geq 0\) and \(Q \geq 0\) and symmetric.

5 Experiments

- **Synthetic data:** We randomly generate 100 datasets and carry out the above algorithms while varying the amount of training data. Figure 1 plots the resulting normalized discounted cumulative gains, defined as
  \[
  \text{DCG}_p = \sum_{i=1}^{p} \frac{2^{r(i)} - 1}{\log_2(i + 1)}
  \]

- **Real data:** Our real data set consists of 30 CT scans of liver lesion. Figure 2(a) gives some sample images. Figure 2(b) plots the NDCG delivered by each algorithm.

References