Fair Neighbor Embedding

February 22, 2024

Seoul National University

Table of Contents

1 Contribution

2 Notation

3 NeRV

Contribution

- Introduce a new fair nonlinear dimension reduction method.
- Low-dimensional embeddings that preserve high-dimensional data neighborhoods without the biased association of protected groups.

Table of Contents

1. Contribution
2. Notation
3. NeRV

Notation

- $\mathcal{X}$: Feature space in $\mathbb{R}^D$
- $\mathcal{Y}$: Embedding space of feature space in $\mathbb{R}^d$
- $\mathcal{S}$: Set of sensitive attributes

Table of Contents

1 Contribution

2 Notation

3 NeRV

NeRV

- The Neighbor Retrieval Visualizer is a nonlinear DR method that aims to create low-dimensional data embeddings.
- Given the input data set $\{\mathrm{x}_i\}_{i=1}^{n}, \mathrm{x}_i \in \mathbb{R}^D$, the probability that data point j is picked as a neighbor of point i is

$$p_{ij} = \frac{\exp\left(-\left\|\mathrm{x}_i - \mathrm{x}_j\right\|^2 / \sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\left\|\mathrm{x}_i - \mathrm{x}_k\right\|^2 / \sigma_i^2\right)}.$$

  where σ_i^2 controls falloff of the p_{ij} with respect to distance.
- NeRV outputs an embedding of the points $\{\mathrm{y}_i\}_{i=1}^{n}$, $\mathrm{y}_i \in \mathbb{R}^d$, in a d-dimensional output space, where neighbor probabilities $q_i = \{q_{ij}\}_{j=1,\ldots,N, j \neq i}$ are defined based on output coordinates y_i is

$$q_{ij} = \frac{\exp\left(-\left\|\mathrm{y}_i - \mathrm{y}_j\right\|^2 / \sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\left\|\mathrm{y}_i - \mathrm{y}_k\right\|^2 / \sigma_i^2\right)}.$$

Fairness

- The objective function of NeRV is:

$$C_{\mathrm{NeRV}} = \frac{1}{N} \sum_{i=1}^{N} \left(\lambda D_{KL}\left(p_i, q_i\right) + (1-\lambda) D_{KL}\left(q_i, p_i\right) \right)$$

  where λ is a tradeoff parameter, and D_{KL} is the Kullback-Leibler divergence, $D_{KL}\left(p_i, q_i\right) = \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$.
- Objective function of Conditional NeRV when values of sensitive variables are given,

$$\begin{aligned} C_{\mathrm{CNeRV}} = \sum_i & \tau^{\in} D_B\left(p_i^{\in}, q_i^{\in}\right) + \left(1-\tau^{\in}\right) D_B\left(q_i^{\in}, p_i^{\in}\right) \\ & + \tau^{\notin} D_B\left(p_i^{\notin}, q_i^{\notin}\right) + \left(1-\tau^{\notin}\right) D_B\left(q_i^{\notin}, p_i^{\notin}\right) \end{aligned}$$

  where $D_B\left(p_i^{S_i}, q_i^{S_i}\right) = \sum_{j \in S_i} p_{ij} \log \frac{p_{ij}}{q_{ij}} + q_{ij} - p_{ij}$ is a Bregman divergence and $p_i^{\in} = \{p_{ij}\}_{j \in S_i^{\in}}$ and $q_i^{\in} = \{q_{ij}\}_{j \in S_i^{\epsilon}}$

Fair-NeRV and Fair-t-NeRV

- Define

$$r_{is} = \frac{\sum_{j \neq i} \delta\left(s_j, s\right) \exp\left(-\left\|\mathrm{y}_i - \mathrm{y}_j\right\|^2 / \sigma_i^2\right)}{\sum_{j \neq i} \exp\left(-\left\|\mathrm{y}_i - \mathrm{y}_j\right\|^2 / \sigma_i^2\right)}$$

  and

$$\rho_{is} = \begin{cases} 1 - \omega & \text{if } s = s_i \\ u(s) \cdot \omega / \left(1 - u\left(s_i\right)\right) & \text{otherwise} \end{cases}$$

  where $u\left(s_i\right)$ is the overall proportion of sensitive value s_i in the data and $\omega \in [0, 1]$ is a weight controlling influence of value s_i in the neighborhood.

Fair-NeRV and Fair-t-NeRV

- Term of fair embedding :

$$C_{\text{Fairness}} = \frac{1}{N}\sum_i \left(\gamma D_{KL}\left(\rho_i, r_i\right) + (1-\gamma) D_{KL}\left(r_i, \rho_i\right)\right),$$

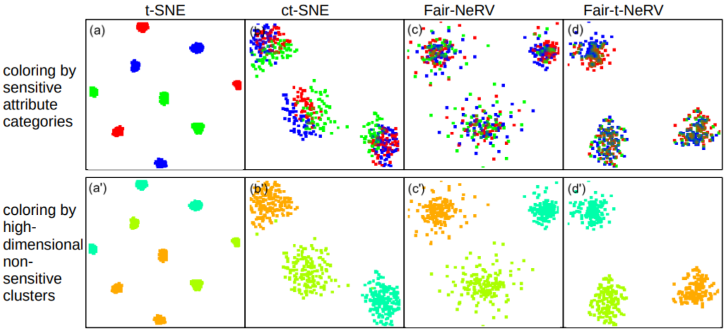
- Final Objective

$$C_{\text{FairNeRV}} = \beta C_{CNeRV} + (1-\beta) C_{\text{Fairness}}$$

Experiment

- The Syn data set is an artificial set of 1000 data points with 5 dimensions.
- The first three dimensions have three multivariate Gaussian clusters, cluster membership considered sensitive information.
- Dimensions 4 and 5 dimension have an independent mixture of three multi-variate Gaussian clusters, considered non-sensitive information

Experiment



t-SNE
ct-SNE
Fair-NeRV
Fair-t-NeRV
coloring by sensitive attribute categories
coloring by high-dimensional non-sensitive clusters
(a)
(b)
(c)
(d)
(a')
(b')
(c')
(d')