Representation learning on gene expression data

CSIC 5011 Final project

Xinwei Shen and Yunfei Yang

The Hong Kong University of Science and Technology

May 19, 2019
Data representation plays an essential role in machine learning.

High-dimensional data often concentrate around a low dimensional manifold.

Data representation methods can help us reduce the dimension and extract useful information when building classifiers or other predictors.
Three levels of representation learning methods:

- **linear dimension reduction methods**: PCA and robust PCA
- **Non-linear dimensionality techniques**: manifold learning
  - Isomap, Locally linear embedding (LLE) and t-distributed stochastic neighbor embedding (tSNE)
- **Deep representation methods**: variational autoencoder (VAE)
Our work

- We explore three levels of representation learning methods:
  - linear: PCA and robust PCA
  - non-linear: manifold learning
  - deep: VAE
- We apply these methods to a real single-cell expression data set.
- We also give detailed analysis and comparison of different approaches.
The Mouse Cortex Cells dataset contains
- 3,005 mouse cortex cells
- 558 genes
- labels for 7 distinct cell types

Cell clustering: to identify the distinct cellular subtypes or states.
- (1) Dimension reduction
- (2) Clustering (k-means, Ward’s hierarchical clustering method)
Data reduction and visualization

(a) MDS  (b) Isomap  (c) tSNE

Figure 1: Visualization of the dataset
Data reduction and visualization

PCA and robust PCA
- Top ten PCs and robust PCs

Figure 2: Cumulative percentage of eigenvalues
Clustering

- Performance measure: adjusted Rand index (ARI)
- PCA + kmeans: 0.4030
- RPCA + kmeans: 0.3817
- This dataset is well preprocessed and is of high quality, the outliers may be removed beforehand.
Clustering

- Manifold learning (Isomap, LLE and tSNE) and VAE + k-means and Ward’s hierarchical clustering

**Figure 3:** ARI of k-means and Ward’s hierarchical clustering with respect to different latent dimensions
Clustering visualization

(a) true labels  
(b) k-means, PCA, dim=10  
(c) k-means, VAE  
(d) k-means, LLE, dim=10  
(e) k-means, tSNE, dim=15  
(f) Ward, Isomap, dim=20

Figure 4: Visualization of some clustering results
Conclusion

- We explore three levels of representation learning methods.
- We apply various methods to single-cell expression analysis.
- We also give detailed analysis and comparison of different approaches as well as different settings of hyperparameters.
- In conclusion, manifold learning and VAE outperform linear dimensionality reduction approaches on this dataset.
Thank you.
References


