CSIC 5011 Mini-Project 1: Dimension Reduction and Classification on Hand-written Digits

CAO Yang, ZENG Wenqi {ycaoau, wzengad}@connect.ust.hk
Department of Mathematics, HKUST

1. Introduction

We first carried out the visualization of dimensionality reduction in order to intuitively see the effects of different dimensionality reduction methods.
Next we use parallel analysis to determine the number of dimensionality by PCA. Then we utilize PCA for dimensionality reduction and feature selection and pass them respectively to subsequent classification tasks to compare their effects

2. Hand-written Digits Dataset

Our dataset contains information of images of hand written digits. Each datapoint consists of an 16 * 16 image with attached label from {0,1,…,9}. There are around 1000 samples for each 10 class, and 70% of them in training set.

Methodology

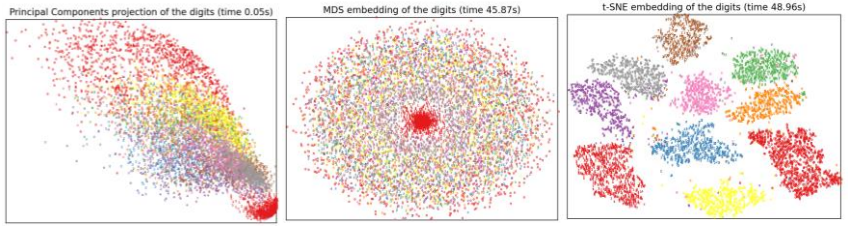
- PCA projects high dimensional data to lower dimensional space. Given by the top right singular vectors of the data matrix, PCA would retain variance of the data as much as possible.
- MDS puts data points into low dimensional space while keeping the distances between data points as much as possible. It can be considered as the transformation given by the top left singular vectors of the data matrix.
- t-SNE is a stochastic way to model the high dimensional data. With high probability, it would put “similar” data points into nearby points and “dissimilar” data points into distant points.

Why?

We want to use PCA to reduce original 256 features in the data and see if there are abundant features in image data. Here we use parallel analysis to guide the choice of components. Another idea for feature selection is to choose a subset of the original features that contains most of the essential information, using the same criteria as PCA. The second method is also known as Principal Feature Analysis (PFA).

3. Visualization

Label “1” might be the easiest one to be distinguished by PCA and MDS. However the data was almost smashed and clusters was completely invisible due to the loss of too much information in the process under the linear dimensionality reduction of PCA. Dimension reduction by t-SNE not only maintains the local structure of the data well, but also maintains the data difference.



PCA MDS t-SNE

4. Classification

	Precision	Recall	f1_score
RandomForests	0.94	0.93	0.94
RandomForests+PCA	0.92	0.91	0.91
RandomForests+PFA	0.94	0.93	0.94
RandomForests+t-SNE	0.91	0.91	0.91

RandomForests outperforms SVM, logistic regression and LDA on hand-written digits. By parallel analysis we keep 26 of PCA components and the number of features selected by PFA is 200. We add 2D t-SNE for comparison considering its good performance in visualization part and use 10-fold cross-validation.

5. Analysis

As we can see, PCA is very fast, while t-SNE would take a lot more time to compute. However, t-SNE gives clear boundaries between classes and make only a few mistakes while PCA and MDS could not since the top 2 singular vectors may not be able to provide enough information.
Adding PCA as a feature selection tool is not promising in this hand-written digits case. While t-SNE performs better than PCA in visualization part (2D), the performance of 26 PCA components in classification process is comparable to 2 t-SNE components.

6. Conclusion and Future work

t-SNE is more capable of exhibiting data-reduced-dimensional clusters than PCA in this case, which may be related to PCA’s attempts to preserve linear structure and t-SNE’s attempts to preserve topology (neighborhood) structure.
The problem with using PCA and PFA as feature reduction is that measurements from all of the original variables are used in the projection to the lower dimensional space where only linear relationships are considered and do not take into account the potential multivariate nature of the image data structure.

7. References

Y.Lu, Feature selection using principal feature analysis. In Proceedings of the 15th ACM international conference on Multimedia, 2007
Y.Yao, CSIC 5011: Topological and Geometric Data Reduction and Visualization, 2020

8. Contribution

Visualization

- CAO Yang

Classification

- ZENG Wenqi