Title: Dimension Reduction and Classification on Hand written Digits

Summary

The authors designed this poster mainly encompass two tasks about the hand written digits dataset, using different dimensionality reduction methods for visualization and implementing multi-classification based on low-dimensional features.

Strengths

This paper possesses the clear intention and also show the comparison results in the two parts. Also, it is innovative to adopt to the Principal Feature Analysis (PFA) approach.

Drawbacks

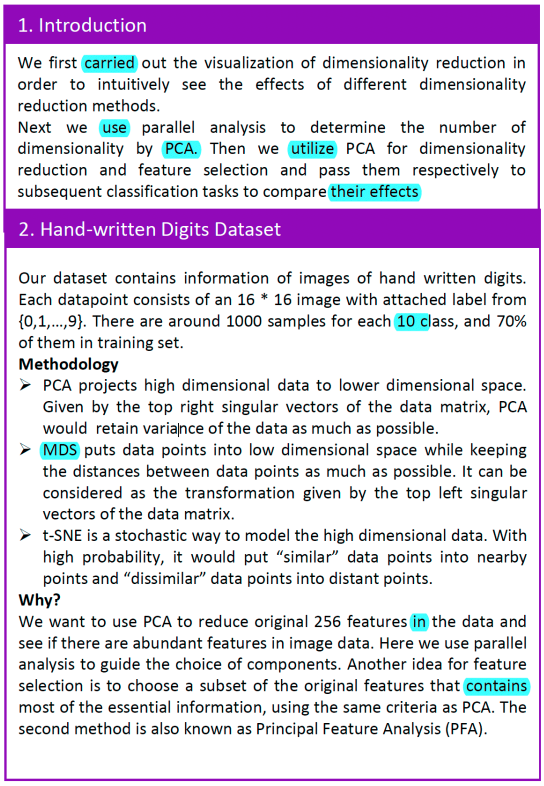
The conclusion part lacks major evidence about classification. The authors ignore the summary in the classification part, only elucidating the analysis of visualization.

Evaluation on Clarity and quality of writing (grade: 3)

Red signs are errors, and blue ones mean corrected result.

For writing part,

1. The poster has too many grammar errors, such as tense error and confusion. In the introduction, the authors use "We first carried out ⋯, Next we use (used), then we utilize (utilized)⋯" . And also, try to avoid singular and plural errors, "Another idea for feature selection is to choose a subset of the original features that contains (contain) most of the essential information" . When PCA and MDS appear for the first time, the shorthand format should not be used. Recommend you use the full academic name.
2. Punctuation missing. "Then we utilize PCA for dimensionality reduction and feature selection and pass them respectively to subsequent classification tasks to compare their effects" .
3. Suggest you delete the "10" in this sentence "There are around 1000 samples for each 10 class" .



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.

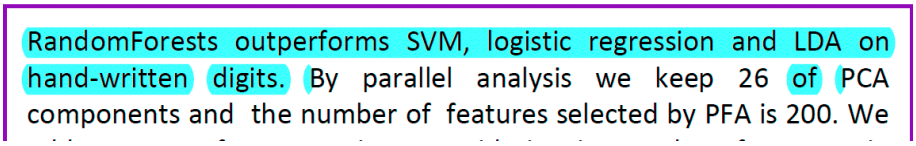
Methodology

- 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).

In terms of references, in the classification part, the authors claim "RandomForests outperforms SVM, logistic regression and LDA on hand written digits." This sentence needs to add one reference to show the compelling evidence. And also, in the figure below, please delete the "of".



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

According to the figure, the figure legends are recommended to be added in the visualization section to prove your conclusion "Label 1 might be the easiest one to be distinguished by PCA and MDS". The figure does not show which color is label 1 and in t-SNE plot the same red color is applied to 2 labels, try to avoid it.

Evaluation on Technical Quality (grade 4)

For the analysis (5) section, the authors' respective is "As we can see, PCA is very fast, while t-SNE would take a lot more time to compute", but the poster doesn't show any time measure to provide this conclusion, so try to show the running time results.

For the conclusion and future work (6) part, the authors partially analyse the results. The work show "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" to highlight the advantages of t-SNE

compared with PCA and PFA, however, in the classification part, the performance of RandomForests with t-SNE is worse than RandomForests+PCA and RandomForests+PFA, so in my view, this work maybe only focus on the result of visualization but neglect the essential results of classification.

Overall rating: 3.5

Confidence on your assessment: 3