ASSIGNMENT 3 ON " ADVANCED DATA SCIENCE ASIIGNMENT 2"		
Student's Code	AIMS African Institute for Mathematical Sciences CAMEROON	Deadline
Gibbs Nwemadji		**03.04.21, 8:30 am**
January 13, 2022		Ac. Year: 2020 - 2021
		Lecturer(s): "Dr. Bubacarr Bah"

Abstract

This project about performing clustering on a PCA datasets in two different ways.One is the performing PCA on the whole datasets and then cluster and second is performing PCA on the sub-data sets and merging them after then cluster.PCA is used as a pre-processing tool and thirdly cluster the six feature sets .The project conclusion was drawn from the dataset which suggested that performing PCA on the sud-data sets and merging them to cluster is the best way.

1 Introduction

Clustering is a unsupervised learning problem.Clustering is mostly used to discover interesting patterns in data.There are two main families of clustering namely direct partitioning and hierarchical clustering.The main difference between these types are with the direct clustering we pre-specify the number of clusters whilst with the hierarchical we do not pre-specify the number of clusters.Examples of direct partitioning is the K-means and the example of the Hierarchical clustering is the Hierarchical Ascendant clustering. There is also mixed clustering which combines K-means and hierarchical clustering .It takes the advantages of the K-means and hierarchical clustering to its advantage.

1.1 Objectives

1. To use PCA as a data pre-processing method on the six feature sets and cluster.
2. To use PCA as a data pre-processing method on the whole dataset and cluster.
3. To compare the performance of above mentioned objectives.

1.2 Datasets

This dataset consists of six features files of handwritten numerals(0-9) extracted from the collection of Dutch utility maps.The handwrtten numerals are classified into 10 classes that is from 0 to 9 .Each feature has 200 patterns per class summing up to a total of 2000 patterns.Within each of the six feature sets the features are different.This sum the entire datasets to a total of 1298000. Below is a more information of the six feature sets with their respective number of features. Corresponding patterns in different feature sets (files) correspond to the same original character.

Data description

1. mfeat-fou: 76 Fourier coefficients of the character shapes

2. mfeat-fac: 216 profile correlations
3. mfeat-kar: 64 Karhunen-Love coefficients
4. mfeat-pix: 240 pixel averages in 2 x 3 windows
5. mfeat-zer: 47 Zernike moments
6. mfeat-mor: 6 morphological features.

2 Methodology

Principal Component Analysis(PCA) is used as for our data processing to reduce the high dimensionality of our data sets and also to visualize the datasets since it is huge.PCA used on the dataset in two different ways.Python is the software programming language used in this analysis.

2.1 Data pre-processing

The datasets were separated by a space this made the loading it in python difficult.We saved the dataset as a csv file and then transform the spaces into a comma using excel before we imported it in python. We labeled the variables in each datasets by creating a row.

PCA on the six feature sets separately

1. We imported all the crucial libraries in python like pandas,numpy,K-means and others.
2. we imported our dataset.
3. we created a different column called $true\_label$ which contains the classes $(0 - 9)$. The first two hundred rows belongs to "0" class followed by the next 200 rows which belongs to "1" class and so on to the last 200 rows which belongs to class "9".This $true\_label$ is created to help us in the measuring the performance of our clustering.
4. For the rest of the pre-processing methods we excluded the $true_labels$this because we do not need it during our pre-processing.
5. we explored the datasets for all the six feature sets separately even though we could not visualize any due to the high dimensionality,we checked for missing data,checked what each column or row is telling us and also checked the variability in the each feature in the six feature sets.
6. There were variations in the data sets for each feature in the six feature sets that is to say during the exploratory analysis of the dataset for mfeat-fou and the other features there were variations in each of them.
7. We standardized each of the six feature sets separately. We standardized the features of each of the six feature sets so that when we apply PCA on the data set it does not give more emphasis to those having higher variances than to those with very low variances.The standardization process centered the values around zero mean and a unit standard deviation.
8. We then performed PCA on each feature of the six feature sets separately.We took a threshold of 95% that is, we want our reduced dataset for a feature to contain at least 95% information from the original dataset of that feature.we repeated the same process for each of the six feature sets.

9. After the applying the PCA and taking at least 95% information from the original dataset, these were our principal components for each of the six feature sets.

    1. mfeat-fou: 58 Fourier coefficients of the character shapes
    2. mfeat-fac: 30 profile correlations
    3. mfeat-kar: 50 Karhunen-Love coefficients
    4. mfeat-pix: 91 pixel averages in 2 x 3 windows
    5. mfeat-zer: 16 Zernike moments
    6. mfeat-mor: 3 morphological features.

PCA on the entire dataset

1. We combined all the six feature sets and had 2000 rows and 649 columns.

2. we explored the dataset and notice variations in the values. so we standardized the datasets to center the values around a mean of zero and a unit standard deviation.

3. we performed PCA on the dataset and took a threshold of 95% information from the original datasets.The 95% threshold is taken so our new dataset (smaller data) has most of the original datasets.

4. we had 154 principal components after performing the PCA on the original dataset.

2.2 Algorithm Building

The algorithm for clustering datasets for PCA on the features and PCA on the entire dataset was the same.

Algorithm method

After pre-processing the dataset, we clustered the datasets using K-means to pre-specify 10 number of clusters.Ten clusters was pre-specified because of the 10 digit classes.We iterated the algorithm 10 times and use a random state of 42.

Merged PCA of the six feature sets

we finally merged the six feature sets we have performed PCA on and repeat the algorithm for clustering as done in algorithm method.

2.3 Results

PCA on the entire dataset The accuracy of the model on the entire dataset was 91.5%.Below is the graphical representation of the 10 clusters for the first two principal components.

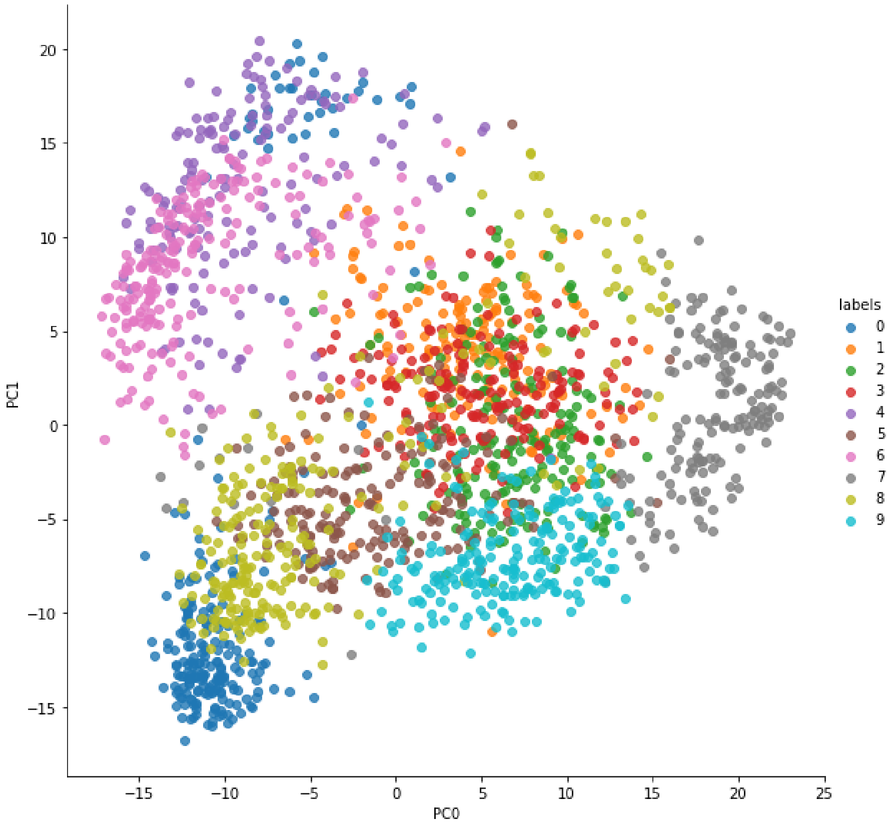


Figure 1: Graphical represenation of clusters

Merged PCA on the six feature sets The accuracy of the model on six feature sets was 93.6%.Below is the graphical representation of the 10 clusters for the first two principal components.

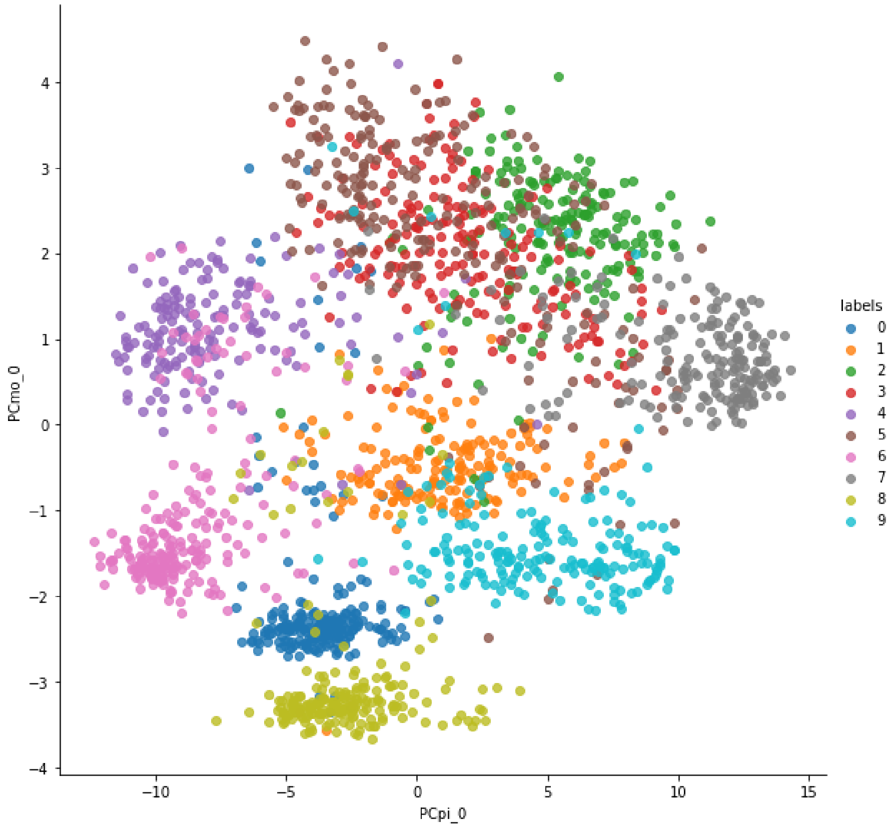
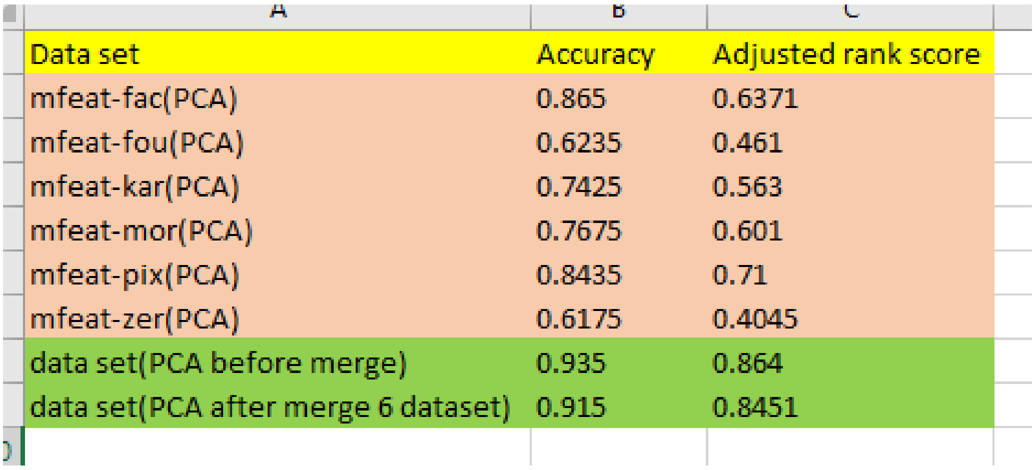


Figure 2: Graphical representation of clusters

PCA on the six features separately The table below gives the accuracy of the six feature sets when clustered.The graphical representaions are found on the python notebooks .



Data set	Accuracy	Adjusted rank score
mfeat-fac(PCA)	0.865	0.6371
mfeat-fou(PCA)	0.6235	0.461
mfeat-kar(PCA)	0.7425	0.563
mfeat-mor(PCA)	0.7675	0.601
mfeat-pix(PCA)	0.8435	0.71
mfeat-zer(PCA)	0.6175	0.4045
data set(PCA before merge)	0.935	0.864
data set(PCA after merge 6 dataset)	0.915	0.8451

Figure 3

2.4 Discussion

When merged the six feature sets as a dataset and performed PCA on the data we had 154 principal components and the K-means performance was 91.5%.
When we performed PCA on the each feature of the six feature sets, each feature in the six feature sets had different number of principal components.Merging all the principal components we had 248 principal components for the dataset in total and performed K-means.The K-means performance on this dataset was 93.6%.

When we performed PCA on the six feature sets, each feature gave an accuracy which was less than the accuracy of the entire datset and the merged PCA on six feature files.

When we combined the six feature sets the correlation between features(variables) reduced and this reduced the performance of the PCA.However, when performed PCA each feature in the six feature sets most of the information in original datasets were in the principal components since there is a correlation in each feature dataset. Clustering datasets performed well because all the principal components contain most of the information of their original datasets.

3 Conclusion

Based on the results of this project, we can conclude that performing PCA on a whole dataset and clustering is not an efficient way to cluster a datasets which has different features to more precise an uncorrelated features.However, performing PCA on correlated datasets before clustering is a best method.Also performing PCA on the different features and clustering them is not best method.We can conclude that performing PCA on large datasets before clustering is a good method to visualize your data and also to reduce a high dimensional dataset..

4 Reference

1. http://archive.ics.uci.edu/ml/datasets/multiple+features