Data Assessment After Dimensionality Reduction

ABSTRACT

PURPOSE: Implement principal components analysis (PCA) using SVD of zero-mean data on raw and standardized data to create effective cluster of data.

METHODS: Measured using Davies-Bouldin (DB) index, principle components analysis is applied to original and standardized dataset.

RESULTS: The plot with PCA reduction on standardized data provides the most clarity between clusters, followed by 13D data clustering, and PCA on unstandardized data.

CONCLUSION: PCA can be used to reduce the dimensionality of data. This process can be optimized by performing modification on the data itself to enhance result. In this assignment, standardized data is proven to be effective in improving clustering results.

Word count: 103

INTRODUCTION

The objective of this study is to determine the effectiveness of clustering of data on original 13D data, and principle component analysis (PCA) on raw and standardized data.

PCA solves the problem of finding the principle ways that the variables differ from the mean. It is a statistical process which is widely used for dimensionality reduction. The first n-score of PCA is an optimal approximation of a n-D vector space of the zero mean data. [2] To find the principle components of a dataset, the singular vector decomposition (SVD) of its zero mean matrix is calculated. For any data matrix A, the zero mean data matrix M can be found by subtracting from each column the mean value of that column.[1] The score vector i of PCA can be found by finding the product of the zero mean matrix M and the loading vectors $\overrightarrow{v_i}$, that is $\overrightarrow{z_i} = M\,\overrightarrow{v_i}$. Score vectors capture the most significant variance within the dataset for clustering and visualization.

One numerical score of a clustering result is the Davies-Boudlin index, a simple method to evaluate labeling algorithm. This method measures the sum of ratios of the scatter of a pair of clusters to the distance between the centroids of the two clusters. [3] The smaller the value of DB index, the more effective the clustering appears.

The study focuses on assessing the effect of dimension reduction approach along with data preprocessing to enhance the result of data clustering.

METHODS

This study finds the first two score vectors of PCA on the original dataset to create a clustering of the dataset. The approach implements PCA on both raw data and standardized data.

The first given task is to determine the two variables of best clustering from the original dataset. The goal is to determine the pair of data column that bext explain the clustering classification. Preprocessing of data requires loading in the dataset and forgoing the first column with variable titles. Then an yvec vector is created to store the classification labels from one of the three grape types (1,2, or 3). The yvec vector is extracted from the dataframe first row, this row vector is then transposed into a column vector. An Xmat data matrix is then created from the dataframe remaining rows, except the first row, and transposed. A nested for loop is used to store all the column pairs' DB index value. This is achieved by putting each column pair and yvec as the paramenters to the dbindex function provided. This is a symmetric matrix where pair DB values repeat twice due to the nested for loop setup. After a DB index matrix is created, the index for the minimum value is displayed to the console.

Second given task is to reduce the 13D data to 2D data using PCA, using SVD, and score reduction. The zero mean matrix is determined by taking the columns in the Xmat data matrix and subtracts it by the mean of the column. Applying MATLAB built-in function svd() can achieve the right singular matrix V. The first two columns of V represent the optimum representation of the data in 2 vector space, thus on a 2D plot. These two score vectors are then multiplied with the zero mean matrix, resulting in a dimension reduction to 2 dimensions. The two dimensionals matrix and the yvec is then used as parameters for dbindex built-in function to evaluate the effectiveness of clustering.

Final given task is to standardize the data and score the reduction that the standardization provides. A similar approach to the second task is implemented in the solution. The data matrix Xmat is standardized and the zero mean matrix is calculated by taking the Xmat matrix and subtracts it by the mean of each columns. An SVD of this standardized zero matrix is taken to get the right singular matrix V. Then the two standardized score vector then multiply the zero mean to get a standardized dimension reduction matrix. The resulting matrixand yvec are used as parameter to find dbindex evaluation once again on this normalized approach.

The methodology is tested on "wine.csv" dataset from University of California at Irvine.

The evaluation process was performed by comparing the different dbindex score between the test trials to evaluate the effect of PCA and data standardization.

RESULTS

Table 1: Data clustering Davies-Bouldin(DB) index. First column includes test trial methods, middle column indicates DB index values, and final column dusplays the pair of values in the data that provide the "best" dimensionality reduction.

Test	DB Index	Variables
Data Columns	0.7875	[1 7]
Raw PCA Scores	1.5148	
Standardized PCA	0.6392	

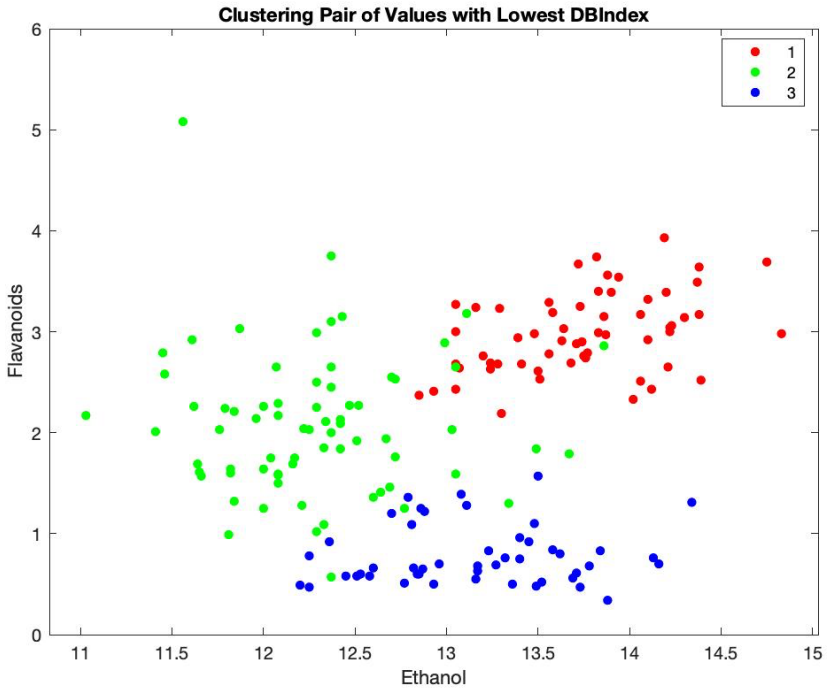

Figure 1: Plot of the orignal data clustering with 13 dimensions. The figure demonstrates clustering into 3 groups of grape type (1,2,3) from the column data [1 7] from the original dataset. The Davies-Bouldin index of this clustering is 0.7875

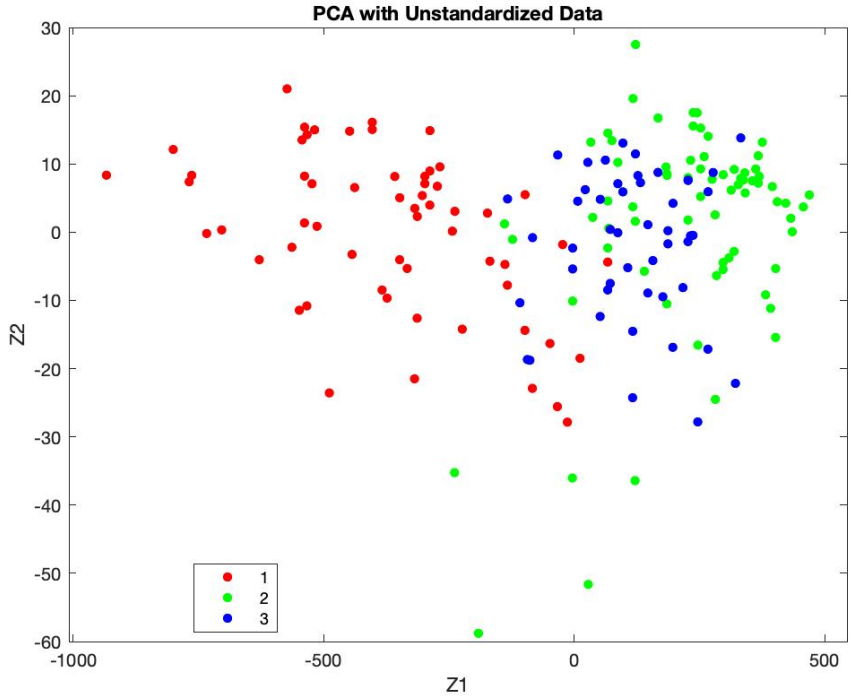

Figure 2: Plot of the PCA on original data. The figure demonstrates clustering into 3 groups of grape type (1,2,3). The Davies-Bouldin index of this clustering is 1.5148

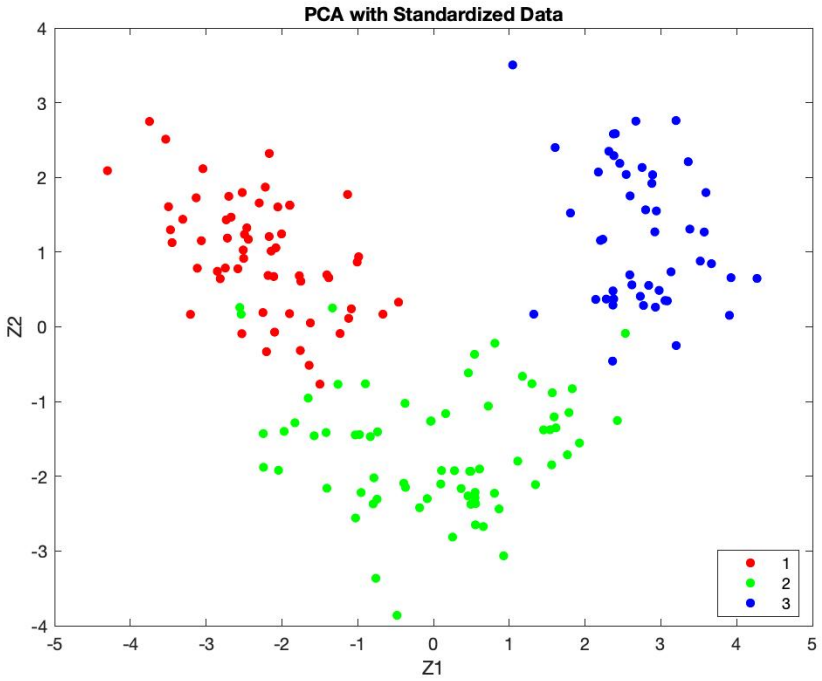

Figure 3: Plot of the PCA on standardized data. The figure demonstates clustering into 3 groups of grape type (1,2,3). The Davies-Bouldin index of this clustering is 0.6392

DISCUSSION

The most effective clustering approach is from the clustering of the standardized PCA dataset. This is shown through the distinction between clusters observed in Figure 3. On the other hand, the plot with PCA on unstandardized data has the worst score out of the three. An explanation for this might be because PCA calculates a new projection for the dataset. The newly created axis are based on the standard deviation of the variables. Thus, a variable with higher standard deviation will have a higher weight for the calculation of axis than a variable with low standard variation. [4] After data is normalized, the variables will have the same standard deviation, thus all having the same weight. This allows the PCA to calculate relevant axis. The idea is that capturing the first two score vectors using unstandardized data creates sparse and intertwined clusters as the PCA is only picking up the highest variance features from the dataset.

PCA can successfully reduces the dimensionality of dataset into variance of features. However, the efficiency of the algorithm is largely dependent on the preoprocessing of data and the choice to either normalize the data or not. With general case, it's best to standardized the data to avoid high variance PCA vector score.

REFERENCES

[1] Ellis Randy E. Class 17: PCA – Principle Components Analysis [unpublished lecture note]. CISC 271: Linear Data Analysis, Queen's University; lecture given 2021 Feb 2021

[2] Ellis Randy E. Class 20: PCA Revisited – Scattered Matrix and Dimensionality Reduction. [unpublished lecture note]. CISC 271: Linear Data Analysis, Queen's University; lecture given 2021 March 2021

[3] Ellis Randy E. Winter 2021 Assignment #3: Assessment of Data After Dimensionality Reduction. CISC 271: Linear Data Analysis, Queen's University; 2021 March
[4] Sautot, Lucile. 2013. Re: Is it necessary to normalize data before performing principle component analysis?. Retrieved from: https://www.researchgate.net/post/Is-it-necessary-to-normalize-data-before-performing-principle-component-analysis/5231b351d4c1186a263d77a7/citation/download.