Data Clustering using Principal Component Analysis and the Davies-Bouldin Index

Abstract

Purpose: Proof of concept for the use of principal component analysis (PCA) to reduce and cluster sample data into data that is easier comprehend, and scoring each algorithm with the Davies-Bouldin Index

Methods: Given a sample data set provided by the instructor, the data was formatted, and was clustered on 3 differet clustering algorithm. Each algorithm was then scored to determine how well the data was clustered.

Results: The clustering algorithm that produced the best clustering (lowest DB index) was standardizing the zero-sum matrix of the independent variables and taking the PCA of the standardized matrix. The next best was the pair of values in the data that provided the lowest DB index followed by the PCA of the raw data.

Conclusions: There a number of algorithms to reduce and cluster data and all 3 metrics were able to reduce and cluster the sample data, but the method that was able to cluster the data the tightest was the standardized 2D PCA data.

Introductions

The purpose of this report is to use the concept of Principal Component Analysis (PCA) and demonstrate its ability to reduce and cluster sample data.

Often when dealing with a data set, there are many variables that describe the data that has been collected. In the data set “wine.csv”, there are 13 independent variables that describe the information we have. PCA is an attempt to reduce the data that has collected to a smaller number of variables, all while keeping as much relevant information as possible. It is based on the idea that not most of the variance comes from a small number of eigenvalues, and we can use these small number of eigenvalues and their corresponding eigenvectors to describe the data in a more understandable way. DB index is a measurement of a ratio between in cluster and between clusters. The optimal clustering solution has the lowest DB index value.

What this paper attempts to achieve is to provide an answer to the question of what clustering algorithm is able to best reduce and cluster the sample data, and potentially future data. This will be done by graphing the data provided by each algorithm on a scatter plot, and find its DB index to “score” the clustering ability of each algorithm

Methods

The algorithm starts by the reading in just the independent data vectors of the sample data which is in the form of a .csv file. The data dependent readings are ordered, so they can be generated. The independent data is then transposed to fit MATLABs format for data matrices. This produces the data matrix that can used to solve the remaining problems. The first question that needs to be answered is what two columns produce the lowest DB index. Using nested for loops, all n^2 possible pairs are tested, and the pair that produced the lowest DB index are outputted as ndx_lo. The next step is to calculate the reduced the data to 2D from 13D and score the resulting clusters. The way this is done is by converting the data matrix to a zero-mean matrix by going through the matrix column wise and subtracting each column by the mean of the column. From this the MATLAB function svd is used to get the SVD of the zero-mean matrix. The PCA of the zero mean matrix is found by multiplying the zero-mean matrix by the two score vectors from the right singular matrix (V). The clustering is then scored by calculating the DB index of this resulting score matrix. Finally, the score matrix produced in the last step is standardized using MATLAB's zscore fucntion, to generate a standardized matrix that has zero mean, and a unit length of 1. Taking the SVD of this matrix, the standardized matrix is multiplied by the two score vectors from the new right singular matrix (V). Again, the clustering produced by this algorithm is scored using DB index. The scores of each of these algorithms are printed to console, and the scatter they produce are generated the best columns/scoring matrices, along with the cluster identifiers.

Testing was done using the sample "wine" csv file as provided by the instructor. It was used to test each part of the function. The function was developed piecewise, making sure each section could provide a valid output before moving onto the next section. Unfortunately, there was no reference to test to output against.

Each clustering algorithm is plotted on a scatter plot according to the variance from each scoring vector (X and Y axises). The clustering score given by the DB index for each algorithm is printed to console as well to give a numerical representation to the data.

Results

Table 1: The resulting DB indexes for each of the 3 clustering algorithms. The 1st and 2nd columns refer to the algorithm and its corresponding index, and the 3rd column refers to the indexes of the data vectors that provide the best numerical clustering.

Test	DB Index	Indexes
Data Columns	0.7875	1,7
Raw PCA	1.5148	
Standardized PCA	0.6392	

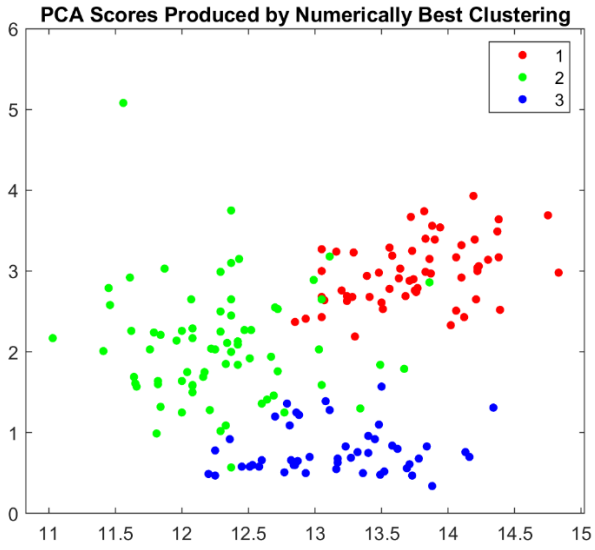

Figure 1: Scatter Plot representing 2D reduction and clustering using numerically best DB index

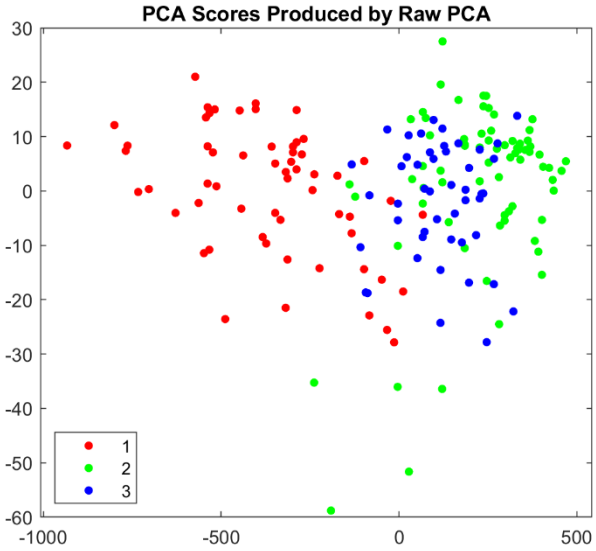

Figure 2: Scatter plot representing 2D reduction and clustering using PCA of the data matrix.

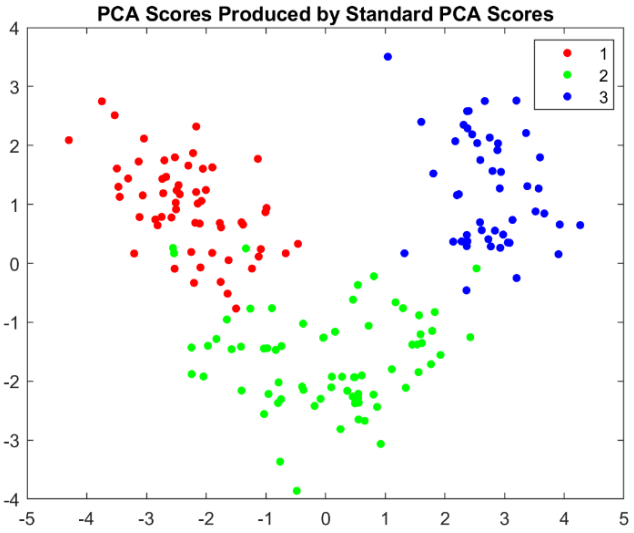

Figure 3: Scatter plot representing 2D reduction and clustering using PCA of the standardized score matrix.

Discussion

The algorithm that had the lowest DB index and appears to have the strongest clustering in the scatter plot was using PCA of the standardized score matrix. The second-best clustering, according to the DB index, is PCA scored by numerically optimal data columns. The worst, at least in this limited testing environment, was PCA by raw data, which did not appear to cluster all that well.

In all 3, group 3 appeared to have the tightest spread, while group 2 has the largest spread. This may be because group 2 has the most samples, and group 3 has the least, but it maybe the nature of the data itself.

Standardizing the score matrix and doing PCA on it was able to reduce the variance 10-fold on the y axis and 100-fold on the x axis. It also reduced the number of outliers produced. Standardizing the data also allowed for separation and clarity between group 2 and group 3. In figure 2, groups 2 and 3 are overlapping, whereas in figure 3 they are distinct from each other. Nonstandard data (group 2) is affected by the fact that variance depends on the units of the data it is based on, and the primary components will change if the units of measurements of one or more of the variables change. Standardizing the data takes these changes of scales out (as they are all based on linear transformations of the standard data set) and allows for the pure variance to be found.

Interestingly, the only plot with no negative values is the data columns, but that may be because there are no negative values in the data matrix.

In conclusion, all 3 algorithms. were able to reduce and cluster the sample data, but the method that was able to cluster the data the tightest was the standardized 2D PCA data.

References

1. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci [Internet]. 2016 Apr [Citied 2020 Feb 24];374(2065). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792409/ DOI: 10.1098/rsta.2015.0202