CENG 414
Introduction to Data Mining
Fall 2019-2020
THE 2

Question 3

1- Using the correlation matrix, which attributes have the highest positive correlation?

Scatter_Ratio and Scaled_Variance_Along_Minor_Axis attributes have the highest positive correlation. Their value in the correlation matrix is 1.

2- Using the correlation matrix, which attributes have the highest negative correlation?

Elongatedness and Scatter_Ratio has the highest negative correlation. Their value in the correlation matrix is -0.97.

3- Below, ”Eigenvalue”, ”Proportion” and ”Cumulative” titles as well as the corresponding numerical values are given. What do they mean? Briefly explain.

Principal component analysis(PCA) is a technique in order to get a less number of variables called principal components from a number of correlated variables. This technique is applied for feature extraction. Feature extraction means that removing the least significant variables in the meanwhile holding the most important parts of variables. This variable removal can be done by combining the important input variables.

Covariance matrix is the measure for each feature that how much it associated with others. Weka gives this matrix for us, but it can be found by multiplying feature matrix and it’s transpose matrix.

Eigenvector is the direction of the special line. This line gives the maximum variance of the dataset in the scatter plot.

Eigenvalue is the relative importance of these directions(eigenvectors).

In the proportion column, there is a value between 0 and 1 which gives the ratio of (eigenvalue/18). For example, the eigenvalue for the first one is 9.42814. By dividing this number to 18, 0.5237855555555556 is obtained. This is the same value with the first value of proportion column.

In the cumulative column, there is number between 0 and 1 as well. This number tells us the value for previous cumulative value + current proportion value.

4- Discuss the results of this analysis. Do you think this dataset is suitable for feature reduction? Justify your answer.

I think that this data is suitable for feature reduction. There is a really big eigenvalue exist in the eigenvalue column which is 9.42814. That means our data is best fit with line for this eigenvalue's direction or eigenvector and this is the most important feature. Also, there are other eigenvalues exist and when we add up these, we get 17.33125.
17.33125/18 = 0.9628472222222223 which is the last value of cumulative column. That means when we select these attributes, our data is represented with .0.96 correctness(this value is greter than expected 0.95). Thus, these 7 features can be extracted and other 11 feature can be reduced.
Selected features are 1,2,3,4,5,6 and 7 based on Weka feature reduction.

Alper KOCAMAN - 2169589