Counterfeit detection in Bank notes with K-means clustering classification.

04/09/2023

Author: Irene S. Diaz.

K Means for detecting forged banknote

Today many Banks in the world are introducing features to improve their banknote's security and actual forgeries can recreate watermarks, metal threads or even holographic features, Modern data science can help Banks identify fake notes more efficiently and take them out of general circulation.

K-means is an unsupervised classification algorithm where data is categorized based on similarity; it's widely used in medical studies and traffic behavior thanks to its effectiveness in dealing with images.

This project proposes a K-means model for classifications of genuine and forged banknotes using extracting features from images by a Wavelet transform tool. “The Wavelets are generally sufficient to analyze data fully”

Aim and purpose of the study

Data exploration

In a study case, a sample is used to estimate the behavior of the general population, the data for this project was provided by: https://www.openml.org/ id=1462. The extracted data were: variance, skewness, kurtosis, and entropy. Of them, in this project, we will work with variance and skewness that from now on will be called V1 and V2. Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean, and variance it' s referred to the pixel variation.

The description of the data obtained is:

- Variable name: V1. variance of wavelet transformed image.Data type: float64.
- Variable name: V2. skewness of wavelet transformed image. .Data type: float64.
- The dimensions of the dataset are: 1372 x 2: 1372 instances x 2 columns or features.
- There are positive and negative values in both variables and do not contain null values. Duplicate values exist, they are found, and removed. The values after: (1348, 2).

Statistical values by variables:

- mean are: V1=0.445785 and V2= 1.909039
- standard deviation(std): V1= 2.862906 and V2=5.868600
- correlation matrix between V1 and V2, we can say that both variables are positively correlated
  V1 1.000000 0.272863
  V2 0.272863 1.000000
- Outliers: 14 values for V1, 17 values for V2.

Analyzing the variance of wavelet transformed image (V1) there is an acceptable dispersion for cluster algorithm. The skewness of the wavelet transformed image (V2) can be considered that the values of this variable show a higher percentage of variability than variance(V1).

Justification of the suitability of the dataset for K-means clustering

The statistical analysis shows that it is possible to apply k means, it has 1372 observations and 2 continuous variables. The data type is also necessary since they are numerical continuous data, float type. Also there is a Class (target) field, where 1 is used for presumably genuine and 2 for false and this is crucial to test.

Modeling.

Finding Optimal K.

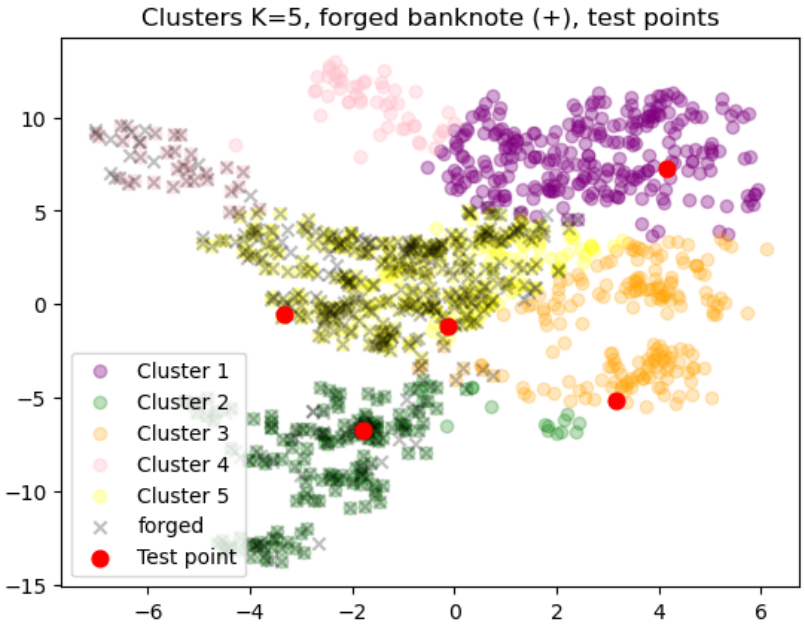
Data was divided in 80% for training and 20 % for testing then normalized with RobustScaler from sklearn. To select the optimal number of clusters according to the dataset two techniques Elbow and Silhouette Both were used both are documented in scikit learn, broadly: **Elbow:** Based on the behavior of the distortion (the sum of the squared distances from each point to its centroid using Euclidean Metrics). **Silhouette** method is based on plotting the density of the data in the clusters found by the model, the score is computed as the difference between the average intra-cluster distance and the mean nearest-cluster distance for each sample, produces a score where 1 is highly dense clusters and -1 is completely wrong clusters. Analyzing the results of both techniques, and relying on the graphs Elbow suggest k=5 and in Silhouettes results, for n_clusters =5 silhouette score is most optimal value: 0.44250951279026096

Clustering

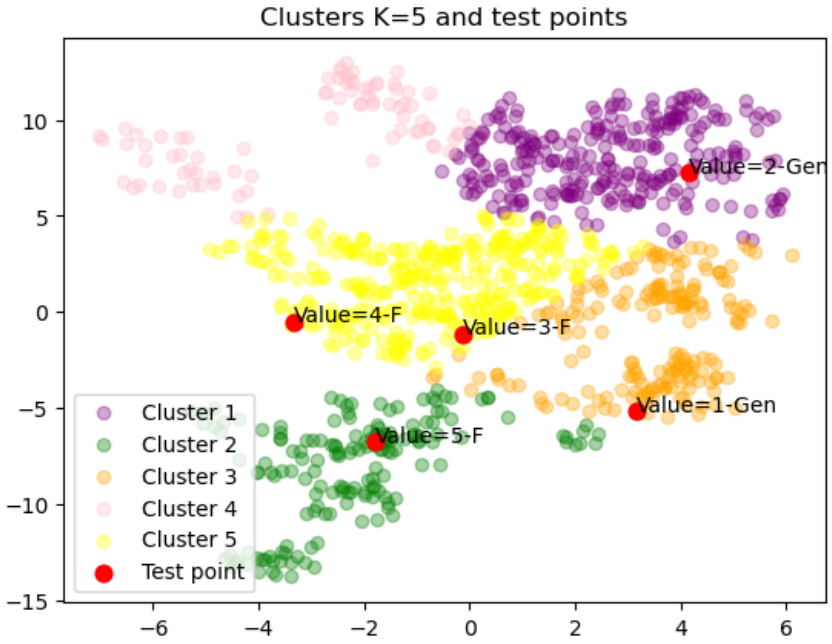
It is important to note that the algorithm converges at iteration 21 in the case of k=5, and becomes stable. After applying the algorithm, the data from the original dataset and the cluster obtained are joined and graphed. Let us rely on the graph to explain the behavior of the data in the five clusters found.

Centers of the model:	Test Values:
[2.64912129, 7.98130978]	1.- (3.1541 -5.171) Target: 0
[-2.1800312 , -7.98638743]	2.- (4.1454 7.257) Target: 0
[3.42403854, -1.09461731]	3 -0.1269 -1.1505 Target: 1
[-3.10634305, 9.48884096]	4. (-3.3458 -0.50491) Target: 1
[-0.72424633, 1.60682876]	5.(-1.7976 -6.7686) Target: 1

Table 1: Data values for centers and test values to test.



Pic: 1. The 5 clusters and forged data values.



Pic: 2. The 5 clusters and test data values.

First: the five clusters are represented on the scatter with different colors and each cluster has been assigned a number.

Second: in addition to the clusters found for the model, the forged data set is graphed, using the marker=+ and the color black. The forgets were obtained from the original dataset with a value of 1.

Third: In clusters 1 and 3 it's observed that all the banknotes are entirely Genuine.

Fourth: The graph also shows some points used to test the model, and it can be noted that in cluster 2, 4,5 it is very difficult to define if the banknote is Genuine or Forge. This means that the prediction percentage for Genuines in the model obtained is not high. Clusters 2 ,4 and 5 therefore represent a weakness of the model created, due to the mixture of Genuine and Forge data.

Recommendations

I. Initial centroids weakness

Is well known that the results of the modeling always depend on the initial values, **Kmeans++** to initialize clusters was used to mitigate this weakness . The values of genuine and false are also plotted to visually interpret which cluster includes the genuine and which the forged, based on the parameters chosen many changes were applied over and over and the results always depend on the initial values, even varied a lot from one number of k to another.

II. Predict using Clusters

Despite the fact that in chosen points to test, the prediction exactly coincides with its value. The accuracy was 100 % Clusters 2, 4, 5 represents a weakness in the model, due to the mixture of Genuine and Forge data.

It is because of this that it is proposed to increase the dimensionality of the model to avoid this problem in cluster three, that is, to choose more variables to train the model.

KMeans is possible to detect counterfeit bills with a trained model but it would be good to investigate with other variables.