Banknote Authentication Project Report
Author: Charlotte Fettes
Date: 13/09/2019

Background:
Counterfeit banknotes are imitation currency produced without legal sanction in. a deliberate attempt to imitate that currency and deceive its recipient. Although counterfeit banknotes make up a small proportion of banknotes – the Bank of England estimates that counterfeit banknotes represent less than 0.02% of banknotes in circulation – they have a number of negative consequences. These include a reduction in the value of real money, and increase in prices (inflation) caused by more money being circulated in the economy, a decrease in the acceptability of paper money, and losses when merchants are not compensated for counterfeit money detected and confiscated.

Project purpose:
Detection of counterfeit banknotes is, therefore, an important task to remove these banknotes from circulation and so counter their ill-effects on society. The purpose of this project is to develop an algorithm that can allocate a banknote into one of two clusters – a forged banknote cluster or a genuine banknote cluster – based on measurements provided on two features extracted from images of that banknote (variance of Wavelet Transformed image, and skewness of Wavelet Transformed image).

The data:
The OpenML dataset on banknote authentication was be utilised. This dataset contains 1372 banknote specimens, with no missing data, and distinguishes between which are genuine (labelled class 1, with a total of 762) and which are forged (labelled class 2, with a total of 610). A number of measures of each banknote specimen was extracted from digitised images of the banknotes and recorded. The features that will be utilised by this project are V1, variance of Wavelet Transformed image, and V2, skewness of Wavelet Transformed image. Figure 1 shows a visualisation of V1 and V2 data using a scatter plot.

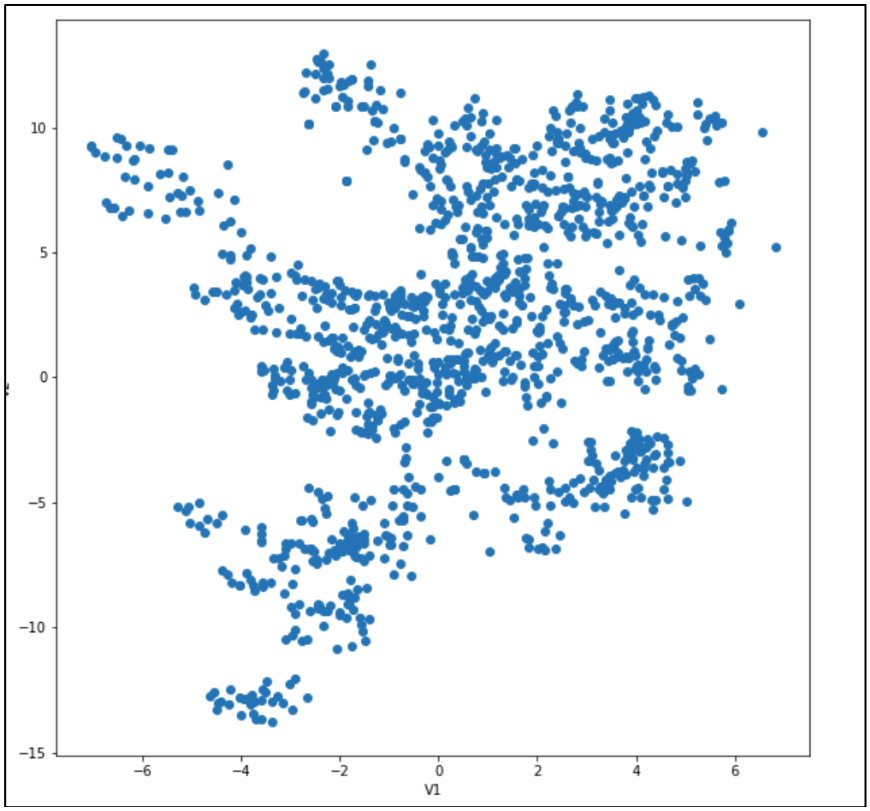

Figure 1: Scatter plot of the raw data for V1 and V2 measures of banknotes

Methods:
Removing the class labels, the V1 and V2 data was subjected to K-means clustering algorithm. A K-means algorithm clusters data points (combinations of V1 and V2 for a single banknote) into homogenous clusters. These clusters are such that those points within the cluster are as similar as possible, and points between clusters are as dissimilar as possible. Therefore, without knowing whether a data point represents a forged or a genuine banknote, K-means will cluster bank notes into groups that are as similar as possible, and so it is expected that genuine and forged banknotes will be grouped into separate clusters.

Limitations of this approach include that the number of clusters (k) must be selected before running. It is difficult to predict the optimal number of clusters (k value). For this project, two clusters were used to represent forged and genuine, however, as explained in the assumptions section, two clusters may not be optimal.

An additional limitation of this approach it that starting conditions can affect the resulting clusters. To combat this issue, after the initial run of the algorithm, this project ran the algorithm 9 more times with randomly selected starting conditions each time to check if the resulting clustering was stable.

The K-means model will also not work if there are too many features, especially if the number of features exceed the number of observations. However, for this project, with 1372 observations and a maximum of 5 features (with only 2 being used for this project), this dimensionality issue should not arise.

A final potential issue is that the reliability of the K-means results can be affected by features being on different scales. For V1 and V2, the ranges, means and standard deviations differ (V2 is approximately double that of V1; see Table 1 below), which may affect the resulting clusters. To remove this issue, the data was standardised for both V1 and V2 to be on similar scales.

Table 1: summary statistics of the V1 and V2 data

	Mean	Standard deviation	Maximum	Minimum	Range
V1	0.4337	2.8417	6.8248	-7.0421	13.8669
V2	1.9224	5.8669	12.9516	-13.7731	26.7247

Assumptions:
Firstly, in this project, it is assumed that all forged banknotes are similar to each other. In the real world, this may not be the case. Many different methods may have been adopted to produce forged banknotes, which may result in forged banknotes with different characteristics. If this is the case, it may not be realistic to group all forged banknotes into a single homogenous group. Multiple 'forged' clusters, and so more than 2 clusters, may be necessary to account for dissimilarities within the group of forged banknotes.

Secondly, it is assumed that the data provided by the study is accurate. As measurements were taken from digitised images of the banknotes, there is always the potential for error. However, it is assumed that the data generated is as accurate as possible.

Results:
The K-means algorithm returned two clusters (green and blue), as illustrated by figure 2 – the stars indicate the location of the central points and cluster boundaries are such that the distance between these central points is minimised (minimising intra-cluster dissimilarities and maximising inter-cluster dissimilarities).

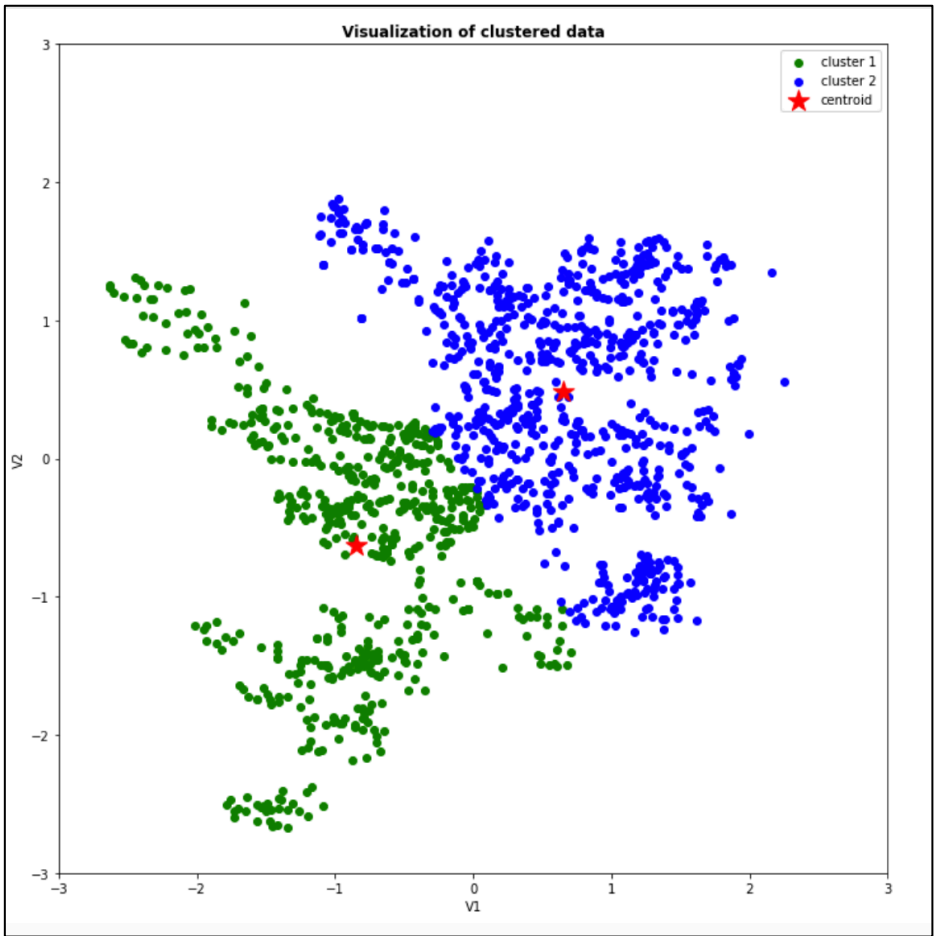

Figure 2: scatter plot of k-means clusters resulting from the first algorithm run

To ensure stability of the clusters, this clustering was repeated with random start points, with the results illustrated in figure 3 below.

The final centroids (stars), regardless of starting position, were the same in all cases, indicating that these clusters are stable (for reference the centroid locations for the standardised data were found to be: (0.65527953, 0.48614465) and (-0.850656, -0.6310923).

In order to check the reliability of the resulting clusters in predicting banknote class – genuine or forged – 10 randomly selected observations from the labelled data were then plotted on the scatter plot along with the unlabelled data. This is shown in figure 4. As can be seen in this figure, all 10 labelled data points are separated into different clusters by class – 3 observations labelled 2 are in the blue cluster, and 7 observations labelled 1 are in the green cluster.

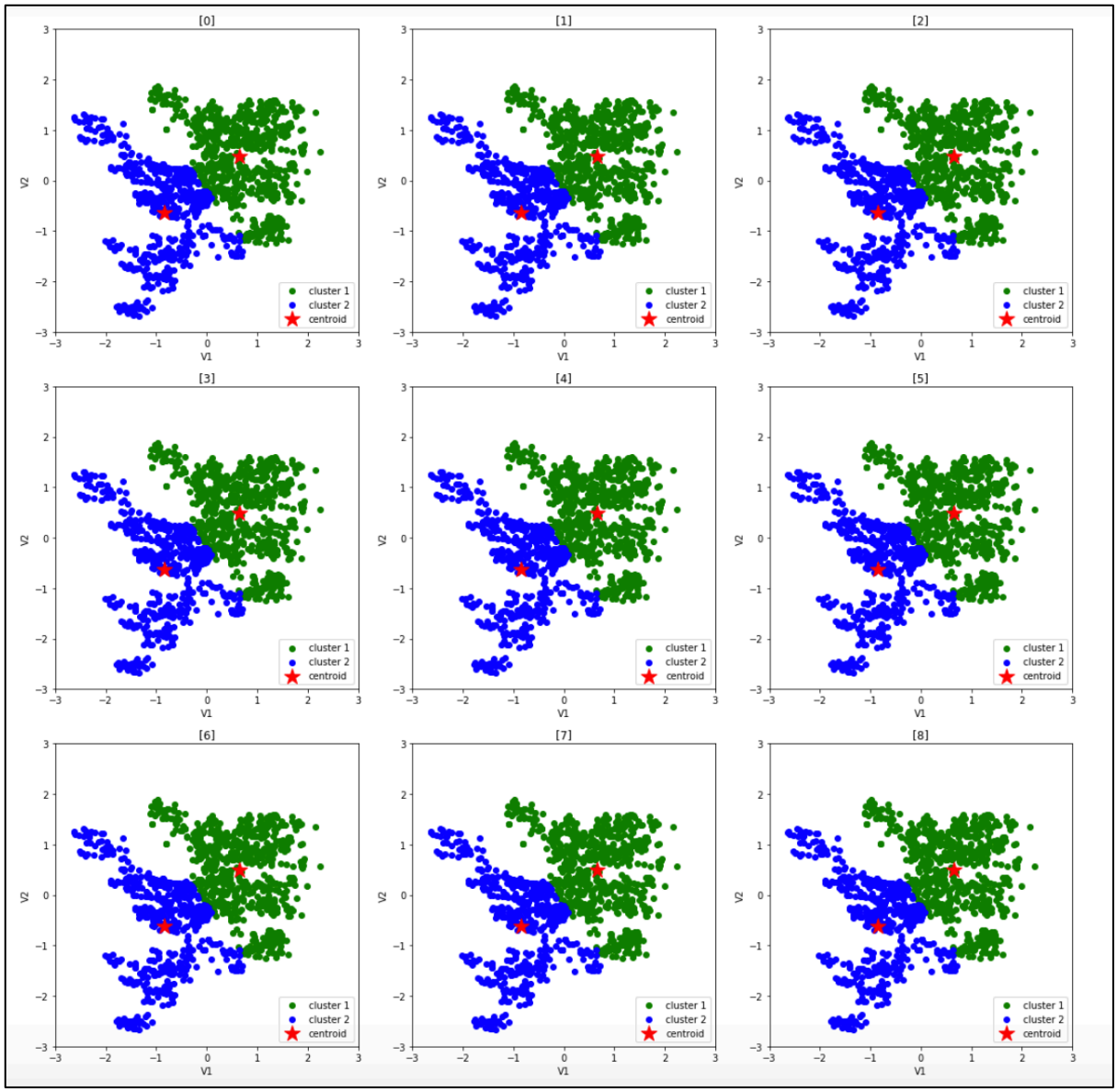

Figure 3: 9 scatter plots of resulting k-means clustered from random initial start positions

Based on these results, it appears that the blue cluster represents genuine banknotes, and the green cluster represents forged banknotes. The results suggest that genuine banknotes have higher V1 and V2 measurements together compared to forged banknotes. However, care must be taken in this interpretation as these V1 and V2 values must occur together – as shown by the scatter plot, although in general the genuine banknotes have higher V1 and high V2 measurements, in some cases forged banknotes can have the same values on one of the measurements, and it is only in combination that they can be recognised as counterfeit.

This data is relevant to banknotes from the currency for which the data were collected from, and the types of forgery used on the counterfeit notes. The results may not be relevant to banknotes of a different currency and if different methods of forgery are used. Furthermore, to truly identify the reliability of the algorithm, the labelled data test should be run multiple times, as it is unlikely that the algorithm will achieve 100% accuracy every time. As can be seen in the scatter plot, there is no clear cluster separation of the data – it looks fairly continuous, and so the algorithm may struggle to identify the banknotes whose V1 and V2 measures are close to this boundary between clusters.

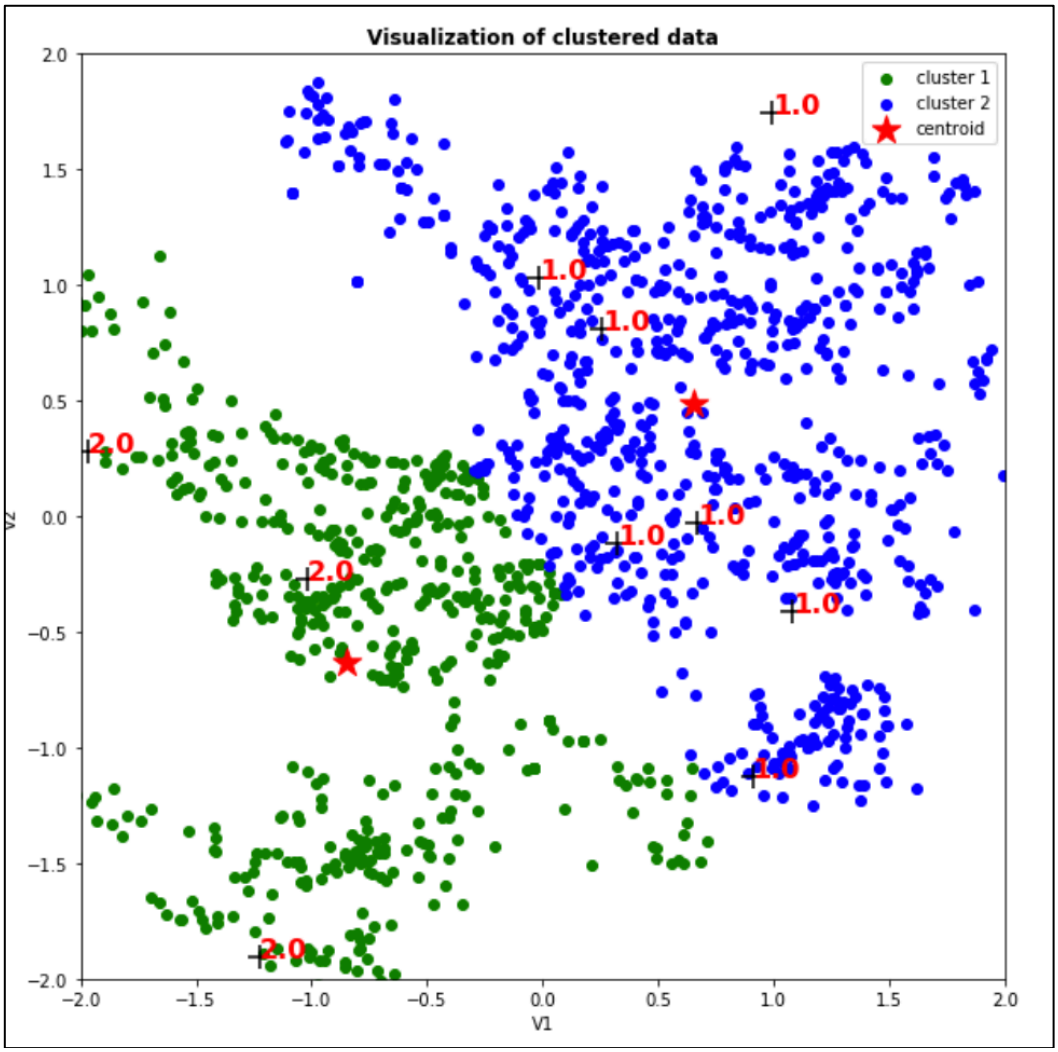

Figure 4: scatter plot of clustered unlabelled and randomly selected labelled data on banknotes

Recommendations:

Measurements of variance of Wavelet Transformed image, and skewness of Wavelet Transformed image are useful features that can be used to detect whether a banknote is genuine or forged; it appears that both low V1 and low V2 measures are present in forged notes, with genuine notes being characterised by higher values of both V1 and V2. This project demonstrated a tentative 100% accuracy in detecting whether a banknote was genuine or forged based on the combination of these measures. However, further testing of the model is recommended to ascertain repeated run reliability.

The results from this project can only be considered relevant to the currency used to generate the data, the types of forgery used in producing the counterfeit notes, how the measurements were taken to produce the data, and the general time that the data was collected (forgery methods may change over time).

Future areas of exploration could be to increase the number of features considered to further ensure accuracy in detection – other features may distinguish the clusters more, giving them a more distinct boundary. In addition, it is recommended to test on other banknotes, where alternative methods of forgery may have been used as this may change the results.

References:

Bank of England, https://www.bankofengland.co.uk/banknotes/counterfeit-banknotes. Accessed 13/09/2019

OpenML Banknote Authentication Data. https://www.openml.org/d/1462. Accessed 13/09/2019