Automatic Detection of Forged Banknotes

By: **Gauden Galea**
Date: **5 May 2019**
Client: **Somesuch Bank (SSB) PLC**

Purpose

Somesuch Bank (SSB) PLC is considering automating the detection of forged banknotes. They have commissioned a study based on a dataset in their possession, composed in roughly equal quantities of real and forged banknotes. The research question is:

> Can a machine learning model be devised that distinguishes real and forged notes in the dataset? If so, what is the predictive power of the model, what proportion of the notes are correctly classified?

Description of the Data

The data consists of 1372 records. Each record contains two observations, `V1` and `V2`, of measurements (called 'wavelets') extracted from digital scans of the banknotes. These wavelets summarise the surface features of the notes capturing aspects of the texture and the dyes used in their production.

It is presumed that forged banknotes will have subtle differences in the textures and dyes and should be detectable from different measurements on the scans. Before an effort was made to build the model, the data was assessed for suitability.

It was discovered that the two sets of observations, `V1` and `V2`, were indeed not comparable in their raw form. `V2` tended to have much higher values than `V1`, and to be much more variable. Both sets of observations were therefore scaled to values within the range from 0 to 1, and their distributions assessed statistically (beyond the scope of this report), and visually. The graphs in Figure 1 below shows the resulting patterns in the data.

The density plot, the graph on the left of Figure 1 shows the shape of the distribution of `V1` and `V2` as a continuous curve between 0 and 1. The graph confirms that the two distributions are centred on the same approximate point, the body of the curves is roughly equal in width and neither of the curves skews to the left or right, but both are roughly symmetrical. The dataset has thus got two comparable sets of observations suitable for further analysis.

Finally, the scatter diagram (the graph on the right of Figure 1) allows a visual assessment of the correlation and clustering of `V1` and `V2`. While there is a large U-shaped concavity on the top left of the scatter plot, the overall shape is that of a rounded cloud, with an internal granular structure. Further analysis below

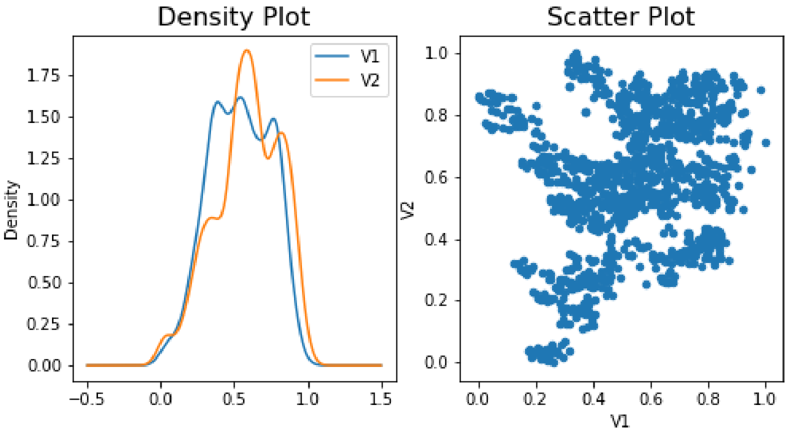

Figure 1: Density and Scatter plot of the two variables

sought to assess this structure and seek to classify the points into the real and forged sets.

Methods

1. The above exploration of the dataset suggested that it would be suitable for **cluster analysis** using the K-Means algorithm. This is a machine learning algorithm that classifies data into groups or clusters of points centred around a common point. In this exercise, since the data is being divided into real and forged banknotes, it was decided to use the algorithm to classify the data points into two groups.
2. The analysis was conducted 1000 times and the clustering of the points was recorded in a database, as well as visualized graphically.
3. The results of the classification were compared with an external independent standard in order to asses the predictive power of the model.

Summary of Results

A cluster analysis was conducted on the dataset of wavelets and the results are displayed graphically in Figure 2. The following points are made:

1. The cluster analysis (using the K-Means package from Scikit-learn[1]) was conducted 1000 times and in all cases the centres of the two clusters converged to the same points (see the coordinates annotated in the diagram).

[1] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

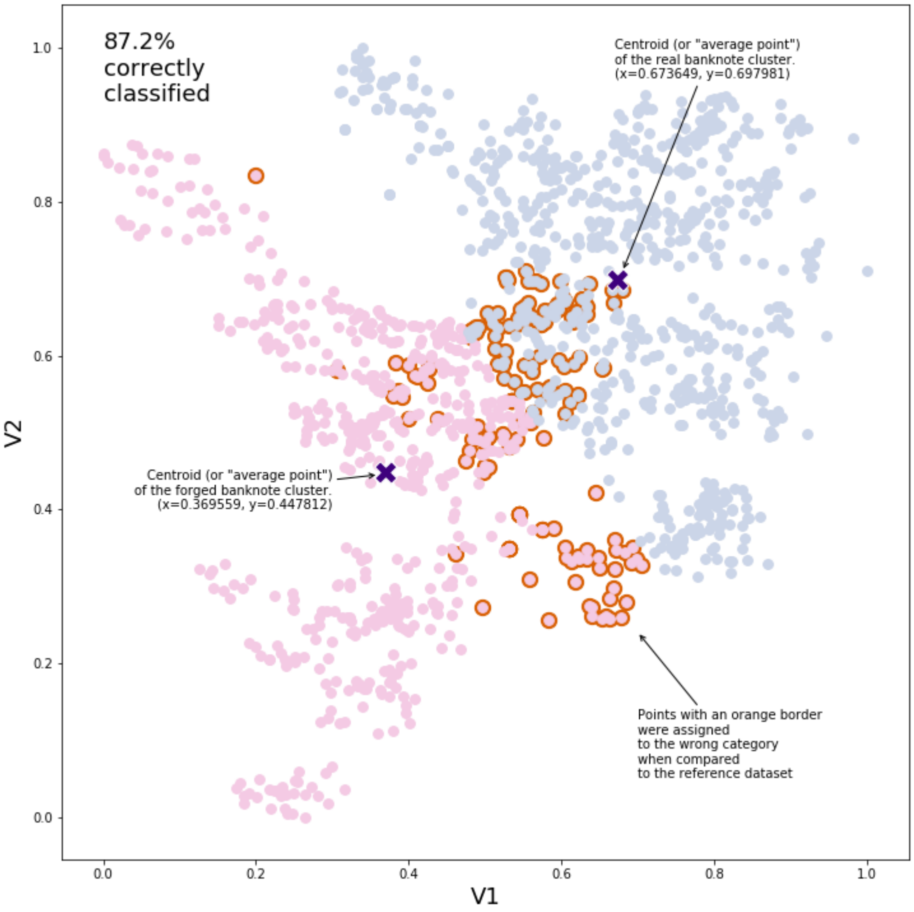

Figure 2: Results of Cluster Analysis of Real and Forged Banknotes

2. The two clusters were labelled ‘fake’ and ‘forged’ based on a comparison with the reference dataset, provided by the OpenML website[2]. The visualization was produced repeatedly and showed that the classification of the data points was stable, whatever the starting point of the analysis.
3. A number of points were misclassified by this analysis in comparison with the reference standard. These erroneous classifications are shown with an orange border. They are largely seen in the border area between the two clusters, and these errors are quite close to the centroid for the real banknotes.
4. The predictive power of this analysis is high, at 87 percent of the points being correctly classified with the reference to the gold standard.

Conclusion and Recommendations

1. *It is possible to develop an automatic means of detection of forged banknotes with high predictive power.* Repeated runs of the analysis on this dataset provided **a stable, consistent classification** of the banknotes into two groups, real and forged. The classification correctly predicted the gold standard in **87.2 percent of cases**.

2. *The current model, despite its high predictive power, is insufficient to deploy in SSB offices.* The current model would incorrectly classify around 13 percent of banknotes leading to loss of revenue for SSB and to distress to honest clients. Three possibilities exist for improving the model:

   a. *Increase the amount of data collected* and re-run the analysis based on a larger dataset. This would probably not increase the accuracy very much but is the easiest to carry out.
   b. *Increase the number of observations made on each banknote.* If more features were scanned and measured on each note, there would be further analysis to conduct. This would probably reduce the misclassifications that happen even in the centre of the clusters in Figure 2.
   c. Finally the results of the model could be validated by other studies, such as comparison with serial number databases and accurate measurement of the size of the notes, which would provide other means of assessing how genuine is any given note. (For examples of this, see the work by J W Lee et al[3])

[2] https://www.openml.org/d/1462 (Accessed 4 May 2019)

[3] Lee, J.W.; Hong, H.G.; Kim, K.W.; Park, K.R. A survey on banknote recognition methods by various sensors.Sensors2017,17, 313