Can Data Science Methodologies Improve the Identification of Forged Banknotes?

A report on analyses of digital photos using K-Means Clustering

Dr N K Thompson

Purpose of the project

The proliferation of forged notes continues to pose a serious problem to banks despite the introduction of significant security features to modern banknotes, such as ultraviolet and holographic features, metal threads, and watermark technology, The intention of this project is to assess whether data science analyses can assist in the identification of forged banknotes by training an algorithm to automatically detect these notes, and thereby help remove them from circulation more rapidly than is currently possible.

Description of the data

The data used from this project were initially generated by the client. This comprised information extracted from digital images of hundreds of banknotes by the mathematical transform tool Wavelets.

These data were supplied as a numerical table that contained 1372 values in each of two columns based on each banknote's 'variance' and 'skewness'. These values can be used to represent the digital 'fingerprint' of each banknote and therefore to compare them with one another.

Metric	Variance (V1)	Skewness (V2)
Count	1372	1372
Max	6.83	12.95
Min	-7.04	-13.77
Mean	0.43	1.92
Standard Deviation	2.84	5.87
Number of outliers	31	41

Table 1. Summary of the data used in this project, illustrating the metrics used during the analyses.

Table 1 presents the main metrics of the data, such as the number of values, their range, and the mean. The metric at the bottom of the table, labelled outliers, is the number of values that lie well outside the general clusters and could potentially be ignored or investigated further, if there were other 'red flags' for these individual notes. However there is not a significant number of these outliers, representing no more than around 5% of the dataset.

Methods: how the data were analysed

These data, being both 2-dimensional and of a relatively moderate size, is considered ideal for testing using a K-Means clustering algorithm in order to achieve the outcome of this project - namely the automated detection of forged banknotes.
In order to analyse these data, it was first necessary to examine the inter-relationship between the two primary data types. To achieve this the data was graphed using a scatter plot as shown in Figure 1 below, which plots the ‘Variance’ on the x axis against the ‘Skewness’ on the y-axis.
As the plot showed a large spread of data it was necessary to assess what outliers might be present, which was achieved graphically by overlaying an ellipse onto the plot in Figure 1. The area of this ellipse, defined by the standard deviation metrics mentioned in the last section (Table 1), visually illustrate the proportion of the data that might be considered outliers (i.e. those points that sit outside the blue polygon).

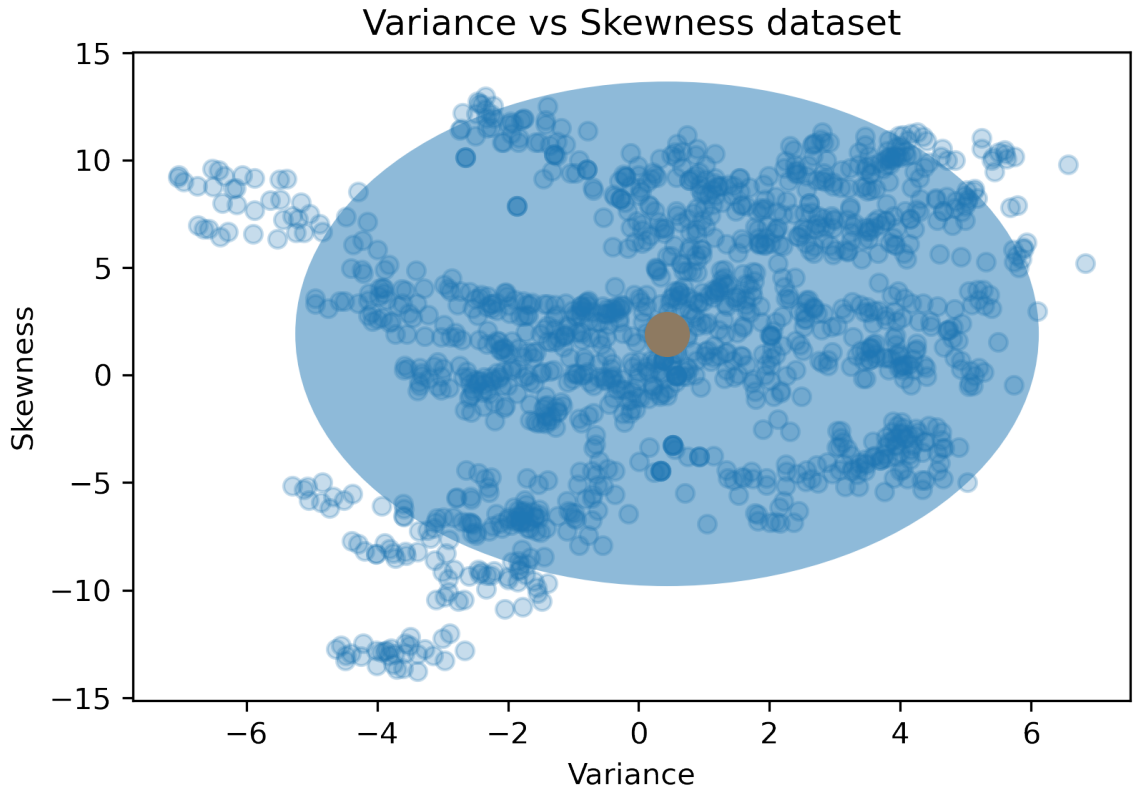

Fig. 1. Scatter plot showing the relationship between the primary dataset, namely ‘Variance' and ‘Skewness'. Overlain on that is the large blue ellipse, centred by a brown dot, which represents all the values within two standard deviations of the mean of the dataset. Those values that lie outside this ellipse can be thought of as outliers from the dataset.

In order to assess the dataset for its ability to predict whether a banknote is genuine or not, the K-Means clustering algorithm was then applied to the data. This process was run using two clusters in order to replicate both potential outcomes, namely those banknotes that are genuine and those that are not. As this dataset consists of only 2-dimensions

(only two different types of data are used) we can expect a good outcome with this particular methodology, as significant numbers of dimensions can increase the chance of an erroneous outcome.
The algorithm was run multiple times in order to assess the variation that occurred from different random start points, which is a function of the way the algorithm processes the data. During this process there was minimal change in the locations of each cluster centre, which suggests that the K-Means algorithm that was run on these data was stable.

Summary of the results

Initial visual investigation of the dataset suggested that there might be a number of clusters within the data, however there was no obvious way to determine visually where the demarkation line between genuine and forged banknotes might lie. The plot shown in Figure 1 served to inform us of the spread of the data, its potential interrelationships, as well as the location of any potential outliers as defined by the overlain ellipse.

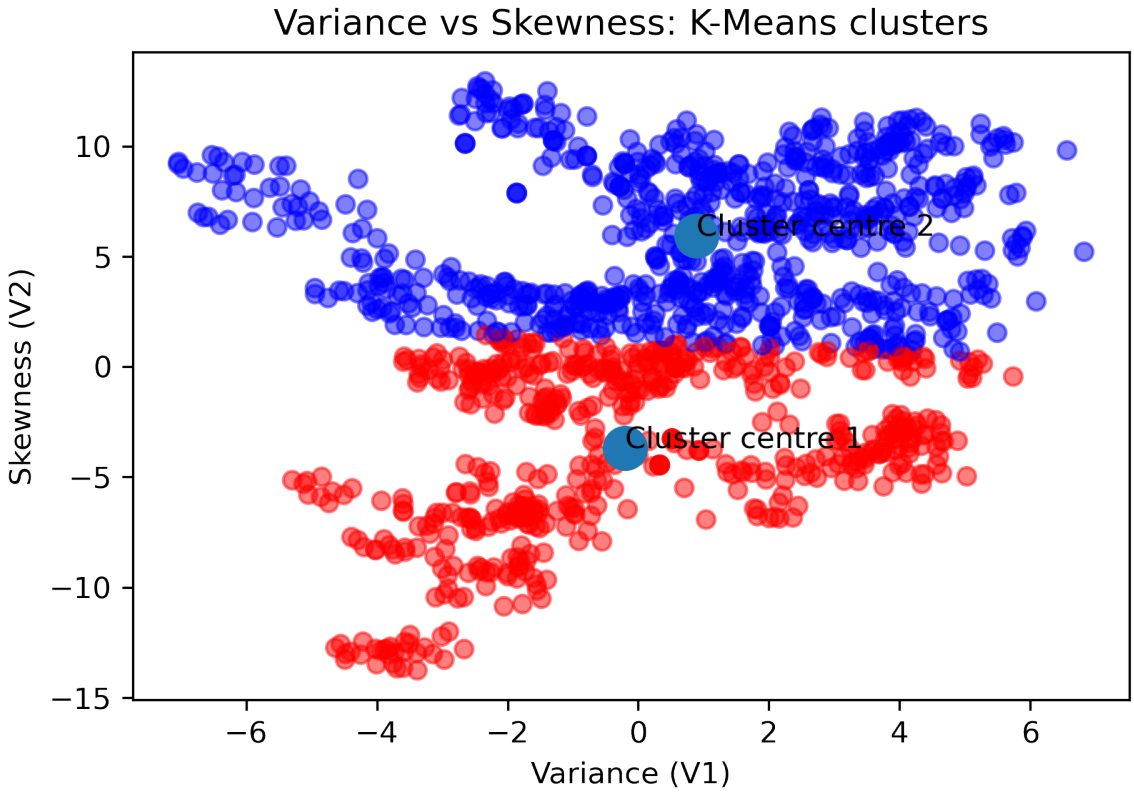

Fig. 2. This scatter plot shows the Variance vs Skewness data used in this analysis, overlain by the two cluster centres derived from the K-Means analysis performed on it. Each cluster is coloured either red or blue, and illustrates the two groups that the algorithm predicts exist within this dataset.

The results of the K-Means clustering algorithm are shown in Figure 2. Two distinct clusters are illustrated in red and blue, with the centres of each cluster labelled on the figure such that there were 780 points in Cluster 1 and 592 points within Cluster 2.
These results were then compared to the outcomes on OpenML, where the assumed correct interpretation of the status of each banknote can be compared with these results.
The comparison showed that 65% of the time the results from this K-Means analysis matched the result from the OpenML dataset. This suggests that this ‘version' of the analysis is still producing a better result than that of a random output (i.e. is better than the 50% outcome that would be expected were the result to be random). Thus is can be considered to be somewhat accurate - although this is based on the assumption that all the results of the OpenML dataset were correct, which may not be the case.

Recommendations to client

The results of this project show that this data science methodology, applied to the problem of determining which banknotes are genuine and which are forged, can certainly improve the speed at which this process can be achieved. However, the process as it currently stands still cannot achieve this result without human input, as the 65% success rate is not nearly reliable enough on its own.
Therefore, the following recommendations are suggested in order to improve this outcome:

- Add further variables to the original dataset derived from Wavelets. This will allow for a more robust result from the algorithm as each banknote will have further layers of unique identifiers, making it easier to determine exactly what sets a genuine note apart from a fake one;
- Increase the size of the dataset used to ‘train’ the algorithm. This will, alongside the first recommendation, further improve the ability to determine the unique characteristics of the forged banknotes and to then differentiate them from the genuine ones;
- Investigate the outlier data points to determine what might be causing these results in the original dataset.