Banknote Authentication: Using K-Means Clustering Algorithm Approach

By Ho Yin Tam

Introduction:

With the emphasis on globalization and international trading, ensuring the authenticity of banknotes is critical to the stability of economics and the confidence of individuals and businesses. Banknote authentication prevents the circulation of counterfeit banknotes and builds the stability of the business environment. In contrast, the circulation of forged currency can result in severe consequences, leading to inflation and loss of confidence in the currency system in a specific country. Also, a reliable banknote authentication allows individuals and business parties to transact with confidence without falling victim to fraudulent transactions. Therefore, in this report, the K-means clustering algorithm is applied to cluster if a banknote is forged or genuine according to the features of variance and skewness of the banknotes.

Properties of data:

The banknote dataset is provided by the course in Coursera 'Foundation of data science: K-means clustering in Python'. It consists of 1327 observations and 2 columns which are numerical data types. The features represent the variance and the skewness of Wavelet transformed image as 'v1' and 'v2' respectively. The task is to classify if a note is forged or genuine depending on the features.

In the process of exploratory data analysis, no missing values are found. Table 1 shows the summary statistics of the data. It is revealed that the minimum and maximum values of the variable 'v1' are -7.04 and 6.82 respectively, while the minimum and maximum values of the variable 'v2' are -13.77 and 12.95 respectively. Figure 1 illustrates the distribution of features of variance and skewness of the banknotes. As the range of the two features is different, feature scaling or standardization is adopted to mitigate the sensitivity of the input features and improve the stability of the machine learning models. Table 2 shows the summary statistics of the standardized data in which the mean and standard deviation are 0 and 1 respectively. Figure 2 depicts the correlation between the variance and the skewness of Wavelet transformed image using a scatterplot after standardization.

	count	mean	std	min	25%	50%	75%	max
V1	1372.0	0.433735	2.842763	-7.0421	-1.7730	0.49618	2.821475	6.8248
V2	1372.0	1.922353	5.869047	-13.7731	-1.7082	2.31965	6.814625	12.9516

Table 1

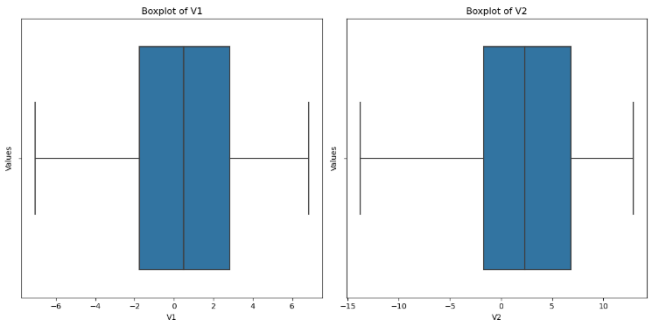

Figure 1

	count	mean	std	min	25%	50%	75%	max
V1	1372.0	0.0	1.000365	-2.630737	-0.776547	0.021974	0.840243	2.249008
V2	1372.0	0.0	1.000365	-2.675252	-0.618819	0.067718	0.833876	1.879908

Table 2

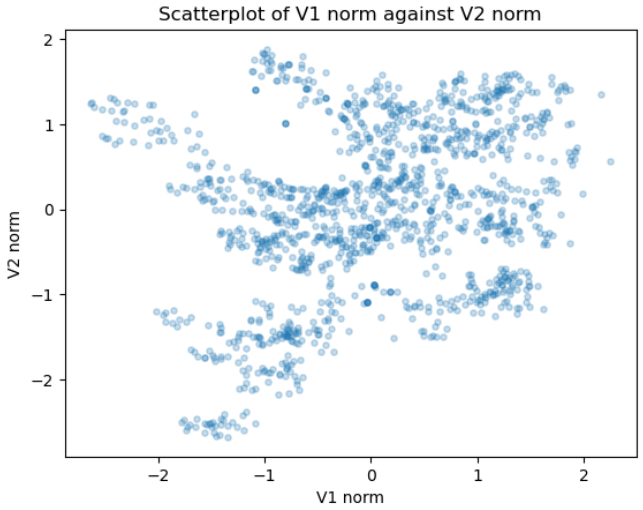

Figure 2

Modelling:

The K-means clustering algorithm is an unsupervised machine learning algorithm and is used to partition the dataset into subgroups. The goal of the algorithm is to group similar observations into the same cluster according to their similarity or nearest mean. In this case, as there are no true labels for the forged or genuine banknotes and the features are numerical data types, the K-Means clustering algorithm is suitable to apply to the banknote dataset. In the modelling process, the cluster of 2 is chosen as the target is to classify if a banknote is fake or genuine.

Analysis:

After several times of re-running of the K-means clustering algorithm, it is revealed that the clusters are stable. Figure 2 illustrates that the K-means clustering algorithm is run for 4 times and the clusters are separated by the purple and yellow colours. It also clearly shows that the centroid of the cluster and the observations in the cluster approximately remain unchanged.

According to the results generated by the K-means clustering algorithm, a banknote with a higher variance (‘v1’) tends to be distinguished as a cluster. In comparison, a lower variance (‘v1’) tends to be classified as another cluster. Hence, a bank can distinguish if a banknote is forged or genuine based on these features. However, as there are no ground true labels, traditional accuracy measures, such as recall, precision, specificity and F1 score, cannot be applied to evaluate the performance of the machine learning model. Though the cluster forming is stable, this does not imply that the K-means clustering model can 100% accurately classify if a banknote is forged or genuine. Employees in the bank still have to pay attention and report to their manager if a suspicious banknote is received.

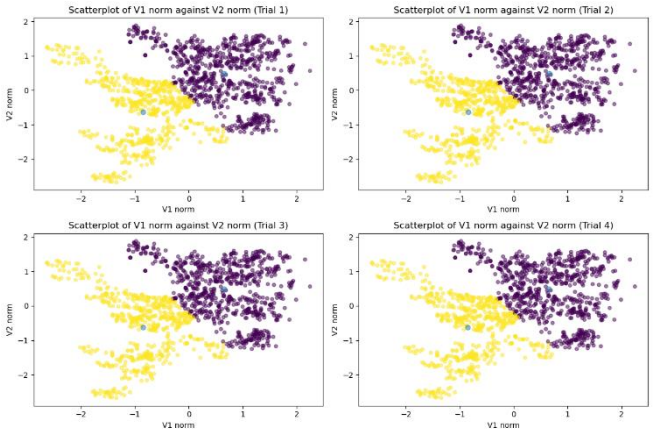

Figure 3

Conclusion:

In conclusion, the K-means clustering machine learning algorithm is suitable for clustering the banknote into two categories, forged banknote and genuine banknote, due to no ground truth labels being provided and the features being numerical variables. In the modelling process, the cluster shows to be stable after being re-run several times as the centroid of the cluster and the observations in the cluster approximately remain unchanged. Individuals have to pay attention even if the cluster is stable as the result does not indicate the model can perfectly and accurately predict and classify if a banknote is fake or real.