Assessing the Possibilities of Automating the Detection of Forged Banknotes

A Langston

1. Problem Definition and Scope

The client, a bank, is considering automating the process of detecting forged banknotes.

The dataset is adapted from an open access dataset available on OpenML. Data were extracted from images that were taken from genuine and forged banknote-like specimens. Within this project, two attributes were examined: V1 (variance) and V2 (skewness).

The project was executed in Jupyter Notebook using Python, both the standard library and the modules Matplotlib, Numpy, Pandas and Seaborn.

2. The Dataset

The data were collected in 2012, and contained 1372 instances, with no missing values. The Wavelet Transform Tool was used to extract the following attributes from the images:

- Variance of Wavelet Transformed image (continuous)
- Skewness of Wavelet Transformed image (continuous)
- Curtosis of Wavelet Transformed image (continuous)
- Entropy of image (continuous)
- Class (integer)

However, only two attributes are necessary for performing K-Means analysis, and therefore only variance and skewness were considered, and will be referred to as V1 (variance) and V2 (skewness).

3. Methods

1. Basic statistical measures were performed to examine V1 and V2, and the presence of any outliers was also studied.
2. V1 was plotted in a scatter plot against V2 to estimate the suitability of the dataset for evaluation using the K-Means clustering algorithm (unsupervised clustering).
3. An initial K-Means clustering, of 2 clusters, was performed.
4. This was repeated 10 times to assess the stability of the K-Means algorithm with this dataset.
5. The outcome and accuracy of the K-Means algorithm as applied to this dataset was assessed.

4. Results Summary

Using the K-Means algorithm to cluster this data into two sets (genuine and forged banknotes) was successful, by the following measures:

1. The algorithm produced stable results when repeated.
2. The accuracy of the model (predicted classification versus actual) was 65.3%.

Therefore, using this model to assess banknotes as either genuine or forged has a stable success rate of 65.3%, without cleaning or otherwise pre-processing the data.

Figure 1 below illustrates the training model in action, based on the plotting of V1 and V2 attributes and applying the K-Means algorithm.

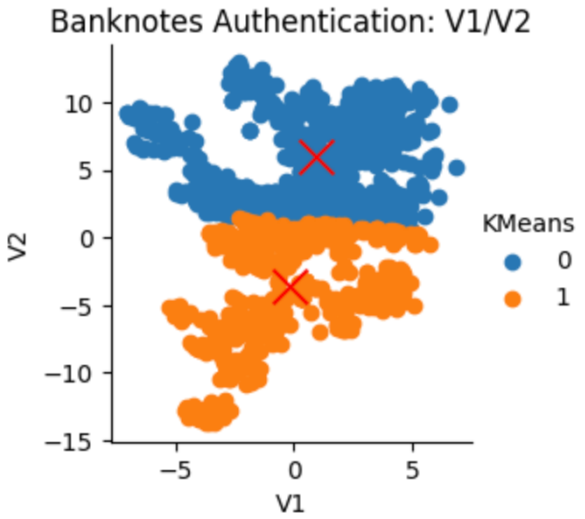

Fig 1: K-Means Clustering Analysis of V1 (variance) and V2 (skew) of Banknote Images to Assess Authenticity

5. Recommendations

The initial assessment of the dataset for suitability of unsupervised detection of forged banknotes indicates that this is a successful way of assessing banknotes.

However, it is predicted that the accuracy level of the model could be improved by:

- Repeated analysis following data cleaning, and/or
- Conducting factor analysis to identify whether V1 and V2 are the most appropriate attributes to use to evaluate and classify banknotes.

While this would delay implementation, it is anticipated that the increased accuracy of the model would reduce the overall time spent in human review of the model outcomes, and therefore any delay would be worthwhile. In the interim, the client can either continue to use the previous non-automated model of assessment, or could use the model developed at this point in this project, and review to capture the 34.7% non-accurate results.