ISSNNo:-2456-2165

Banknote Authentication Analysis Using Python K-Means Clustering

Ragavi E
Final year Engineering Student,
Dept. of Computer Science and Engineering.
Prathyusha Engineering College, Thiruvallur 602 025

Abstract:- The objective is to analyze the given data sets V1 and V2 from the bank_authentication_notes.csv which is taken from open ML datasets, is to identify the forged and real notes using K-Means Clustering Concept forming two distinct clusters of real and forged notes. K-means is easy and simple uses unsupervised learning to solve clustering related problems. It classifies the given datasets to form a group of clusters based on some similarities. The major goal is defining k centers, one for each cluster. The ultimate aim is to use this dataset to train a machine to detect fake notes automatically. However, before implementation, it is important to access if this dataset can sufficiently distinguish forged banknotes from genuine ones. Hence, in this report, with k-mean cluster analysis, unsupervised machine learning, performed on the datasets, we will visualize and outline the results and make according torecommendations.

Keywords:- *K-Means Clustering, unsupervised Learning, Clusters, banknotes.*

I. INTRODUCTION

Forged banknotes are no longer just a problem for merchants, but for banks as well. In recent years, all across the UK and the Eurozone, the service of direct cash deposit at a cash machine has been rolled out. It is of the uttermost importance to find a solution to stop criminal action. Here, we have developed a robust system to identify forged notes by identifying just two features and clustering them in order to predict the forgeries. The dataset is taken from https://www.openml.org/d/1462. This report is consists of data extracted from imagines with Wavelet Transform. The imagines were taken from genuine and forged banknote specimens (n=1372). There are two attributes in this dataset (V1: variance of Wavelet Transformed image and V2 skewness of Wavelet Transformedimage).

In mathematics, a wavelet series is a representation of a square-integral (real-or complex-valued) function by a certain orthonormal series generated by a wavelet. This provides a formal, mathematical definition of an orthonormal wavelet and of the integral wavelet transform.These wavelet of values are defined using variance and skewness are transformed to an image values.

Here the input csv file consists of 1372 instances divided into 4 values as V1, V2, V3, and V4. Here we are using the V1 and V2 values only. The values are classified into two classes 1 and 2.

II. METHODOLOGY

	V1	V2
0	3.621600	8.66610
1	4.545900	8.16740
2	3.866000	2.63830
3	3.456600	9.52280
4	0.329240	-4.45520

Table -1: The values v1 and v2 obtained from the dataset

The Algorithm used here is K-Means Clustering. As it is the simplest method for forming clusters to easily detect the forged and clean banknotes based on the variance and skewness (i.e. V1 and V2). Here, the data is widely diversified, so it is important to normalize the data using

Data normalized = (data - data min () / (data max () - data min ()).

After normalization, the data becomes stable where it comes under a certain range of 0 to 1. Using the data describe, the instances can be described as follows: The data describe function gives the number of instances, mathematical distributions such as mean, standard deviation, minimum value among all the instances, maximum value among the instances ,etc. which are further helpful in the normalization of data.

	V1	V2
count	1372.000000	1372.000000
mean	0.433735	1.922353
std	2.842763	5.869047
min	-7.042100	-13.773100
25%	-1.773000	-1.708200
50%	0.496180	2.319650
75%	2.821475	6.814625
max	6.824800	12.951600

Fig 1:- The properties of data in the dataset

The above Figure -1 depicts the overall description of data in the data set given. We see that the values are extremely diversified ranging from -ve to +ve. So, we cannot form clusters from these values. We must convert them into values which are of a well-defined range in order to groupthem.

The data is normalized using the above formula, which lies in the specific interval makes it easy to plot the data in a scatter plot graph given below:

```
x=data['V1']
y=data['V2']
plt.xlabel('V1')
plt.ylabel('V2')
plt.scatter(x,y)
plt.show()
```

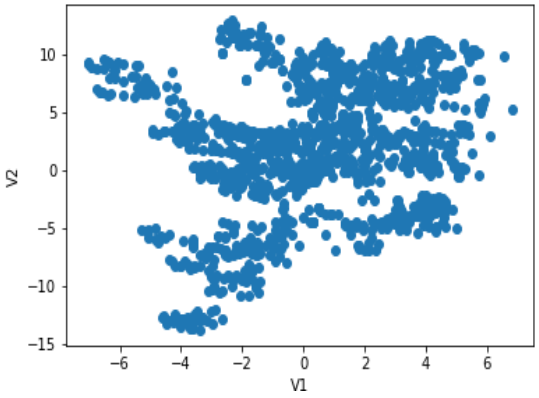

Fig 2:- The initial plotting of V1 and V2 before clustering.

III. MODELLING ANDANALYSIS

The model used here is K-Means clustering, the major objective of K-means is to group similar data points together and discover any similar patterns. The k means clustering will start by finding the k centres to form 'n' clusters. A cluster can be formed by grouping of similar data points aggregated together considering some specific functionalities and similarities. The centroids which are formed is an imaginary or real location at the centre of a cluster identified by k-means clustering forming k-centres. These k-centres allows nearest data points to the nearest clusters forming a 'n' clusterskeepingthedatacentresas smallaspossible.

The *'means'* in the K-means is obtained by finding the centroid, using average distance calculation.

The **Scikit-learn** library is used to import the necessary libraries for the k-means clustering implementation

The following libraries in our project:

- **Pandas**-usedtoreadandwritethecsvfile
- **Numpy** - used to perform mathematicaloperations
- **Matplotlib**-usedtovisualizethedataingraphs
- **Scikit-learn** - used in k-meansclustering

After normalization plot the obtained two clusters as Forged and real bank notes.

The graph forms two clusters at
[[0.65504068 0.48596745]
[-0.85034594 -0.63086227]] Positions.

The two clusters formed by plotting the data before normalization and after normalization of data are:

```
result=np.column_stack((x,y))
km_res=KMeans(n_clusters=2).fit(result)
clusters=km_res.cluster_centers_


plt.scatter(x,y,s=ineq*5,alpha=0.5,color='black')
plt.scatter(clusters[:,0],
            clusters[:,1],s=300,color='red')

plt.show()
```

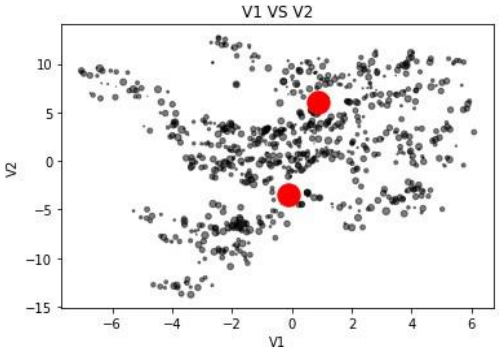

Fig 3:- Cluster before normalizing data

```
plt.xlabel('V1')
plt.ylabel('V2')
plt.title("V1 VS V2")
plt.scatter(data_normalized['V1'], data_normalized['V2'], c=km_res, cmap='summer')
plt.scatter(clusters[:,0], clusters[:,1], s=1000, c='black', alpha=1)

plt.show()
```

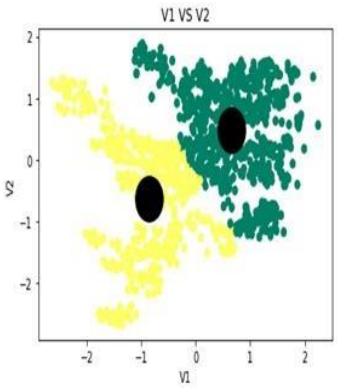

Fig 4:- Visualization of K-Means cluster Analysis

The cluster formed after normalizing data where the yellow region represents forged bank notes and the green region represents real bank notes which are formed by the clusters at centroids represented by black dots**.**

IV. RESULTS AND DISCUSSION

We need to quantify how good our model is for you to implement in your network. Given that we just have two outcomes, real and forged, we proceeded to cluster the data around two points, where each point that belongs to a cluster shares similar features. After calibrating the model, we got the two clusters. After running the k means algorithm a few times, it was found out that the clusters were more or less stable. The cluster is not fully reliable as it might have some tolerance.

With just two parameters, our model only passed 3 forged notes as genuine out of 1,372 banknotes, an error of roughly 0.22.The next step would be to improve the dataset, especially gathering information on genuine bank notes. Going back to the last plot, we see that the data points of features of real bank notes are less disperse than the counterfeit ones.

V. CONCLUSION

The client can able to easily identify the real and fake banknotes from the scatter graph pointed forming two clusters. The analysis was done for the given data set.The data could be better processed if there were more features corresponding to an item in the list. Let us assume that from our plot one data note is termed as fake, but still that data may be free of errors but it could be wrongly classified as fake one. Similarly, a data might contain some errors too but it is classified as real bank note. So, in order to give the right classification, we need a few more parameters.

For example, if there were 4 features instead of two, then we could have analyzed which feature affects more than the other by plotting them against each other which improves the accuracy. I would suggest the client can do additional tests on the banknotes which are classified as forged banknote and to monitor realbanknotes.

REFERENCES

[1]. The 7th python in Science Conference 2007, Aric Hagberg, Daniel Schult and Pieter Swart. Exploring network structure, dynamics, and function using network. Proceedings.
[2]. Journal of Machine Learning Research,2010 .Tom Schaul, Justin Bayer, Daan Wierstra, Yi Sun, Martin Felder, Frank Sehnke, Thomas Rückstieß, and Jurgen schmindhuber, Pybrain.
[3]. Ali Feizollah,Nor Badrul Anuar ,Fair Amalina, Comparative study of k-means and mini batch k-means clustering algorithms in android malware detection using network traffic analysis, published in 2014International Symposium on Biometrics and SecurityTechnologies (ISBAST)
[4]. Sculley D (2010), "Web-scale k-means clustering", Proceedings of the 19th international conference on World Wide Web, Raleigh, North Carolina, USA, pp. 1177-1178.
[5]. Python (2014), "python", Available at: www.python.org (Accessed: 1st March2014).
[6]. Scikit-learn (2013),"sklearn. Cluster. MiniBatchKMeans", Available at the link given below as: http://scikitlearn.org/stable/modules/generated/sklearn. Cluster. MiniBatchKMe ans.html (Accessed: 1st March2014).
[7]. Maulik U, Bandyopadhyay S: Genetic algorithm-based clustering technique. Pattern Recognition.2000, 33: 1455-1456.10.1016/S0031-3203(99)00137-5.