Helix (2021) 11 (3): 1-6

https://doi.org/10.29042/2021-11-3-1-6

Big Data Analytics to Authenticate Bank Notes Using K-Means Clustering

[*1]Mudassir Khan, [2]Mahtab Alam

[*1]Department of Computer Science, College of Science & Arts Tanumah, King Khalid University, Saudi Arabia
mkhankku@gmail.com
[2] Department of Computer Science, Mewar International University, Nigeria

Received: 22nd April 2021, Accepted: 24th May 2021, Published: 30th June 2021

Abstract

The main objective of this work is to use the given dataset (about the information of different variables of the different kind of bank notes that were analyzed) in such a way using an algorithm that will group the large values in the data as efficiently as possible into two groups of values then we can use those groups and make a model that can in future identify a forged note from the real ones. Forged banknotes are a major problem that the bank must tackle. It is very important to find a solution to stop this criminal action. One of the most important steps is to be able to identify which banknote is forged. There are many solutions to identify the forged banknote. In this manuscript, we propose one of the solutions to identify the forged banknotes using one of the machine learning methods called "K-Means Clustering".

Keywords: *Big Data, K-Means Clustering, Bank Notes, Machine Learning, OpenML*

Introduction

Forged banknotes are a major problem that the banks have to tackle. It is very important to find a solution to stop this criminal action. One of the most important steps is to be able to identify which banknote is forged. There are many solutions to identify the forged banknote. In this research article, we propose one of the solutions to identify the forged banknotes using one of the machine learning methods called "K-Means Clustering".

Banknote authentication sticks an important summons for the central banks to remain the robustness of the business system all over the world, and to remaining promise in reliance documents, mostly banknotes. The different approaches encompass the way of digitally processing image to be authenticated the exterior of applicant presented document, which is going to condition of observation incorporates at least part of the security features, the digital processing including executing a decay of the sample image through means of wavelet transform of sample image. The used approach examines currency, the applicability is not easy in the surroundings of Euro banknotes as this currency mandates numerous viewpoints to keep away from the duplications hence different theories on various features and their location should be done.

The world is growing towards technological era and most of the developing countries are focusing on digital transactions. The number of new banks is increasing as per the demand in the increase of number of users. There are many existing techniques available to examine the forged banknotes. The digital currency is also a solution to stop the forged banknotes.

Related Work

Khan, M., et al. (2019) [1] presented big data analytics techniques and Means clustering: 1) K-Means clustering technique 2) based on the partition entropy-based outlier detection technique.

Priyanka B.Mohite, Prof.A.R.Kulkarni, (2016) [2] In this research article author has presented the different Enterprise web Application Using Hadoop MapReduce System. Here the author has tried to analyse the web data and used different approaches to process the data.

In [3], describes the different social media and big data analytics techniques to analyse an iterative procedure of clustering method by using K-means clustering algorithm. Khan, M discussed the different era of big data using different methods and techniques [4][5] and Khan, M. and M. D. Ansari (2019) discussed the security and privacy issues in the present system [6]. [7] Aakashsoor, and Vikas, (2014) has produced an Improved Method for Robust and Efficient Clustering Using EM Algorithm with Gaussian Kernel to detect the forged data.

Sculley D (2010), "Web-scale k-means clustering [8] The authors have given the Enterprise Web Application and given the web scale method to detect to forged banknotes by using k means clustering. The different approaches of Enterprise Web Log Analysis for Security Compliance Using Big Data discussed to draft the existing approaches. [10] Alguliyev, R., (2018) presented their work entitled Weighted Clustering for Anomaly Detection in Big Data. Here the author has given proposed algorithm is applicable to solve the issues occurred during web. The web deriving approaches were discussed to solve the different issues in the pattern discovery network and web browsing.

In [11] Barbara, has given the different Requirements for Clustering Data Stream analysis of big data using K-Means clustering problems was proposed. The different approaches were used by the author Barbara. M et al for the comparative study of web log files using map reduce techniques. In [12] G. Tzortzis and A. Likas has discussed the MinMax K-Means Clustering Algorithm approaches to enhance the existing techniques. The author also proposed method is applied to big data processing techniques. The different approaches used by the authors to solve and process the different problems.

G. Gan. and F. Jiang [13][14] have proposed Initialization of K-modes Clustering using Outlier Detection Techniques and discussed the different categories of k means clustering techniques uses and approaches. The authors have inspected the big data and their existence in the growing world.

Shahdad, S. Y., et al. (2019) [15] proposed a method Routing Protocols for Constraint Devices to check the consistency of data used for processing. Here the author also discussed the different techniques to route the protocols for different constant devices.

Dataset used: The dataset had total of 4 variables that were gathered using wavelet on the images of multiple bank notes 400x400, but for the project we took only 2 values V1 & V2 which represented the Variance and Skewness variations of the images that were analyzed and the number of times each image was analyzed the values of V1 & V2 were recorded and hence the dataset was created. It had 1371 values in it.

	V1	V2
2	3.6216	8.6661
3	4.5459	8.1674
4	3.866	-2.6383
5	3.4566	9.5228
6	0.32924	-4.4552
7	4.3684	9.6718

Fig 1: Values of Variance (V1) and Skewness (V2) Variations used as dataset

Data Description: According to the OpenML.org, the banknote-authentication dataset is the dataset about distinguishing genuine and forged banknotes. Data were extracted from images that were taken from genuine and forged banknote-like specimens. A Wavelet Transform tool was used to extract features from these images.

This data used in this project a simplified version of the dataset available on OpenML. The data contains only 2 columns: the variance of Wavelet Transformed image (V1) and the skewness of Wavelet Transformed image (V2). The descriptive statistics of the data is shown below. The number of samples is 1,372.

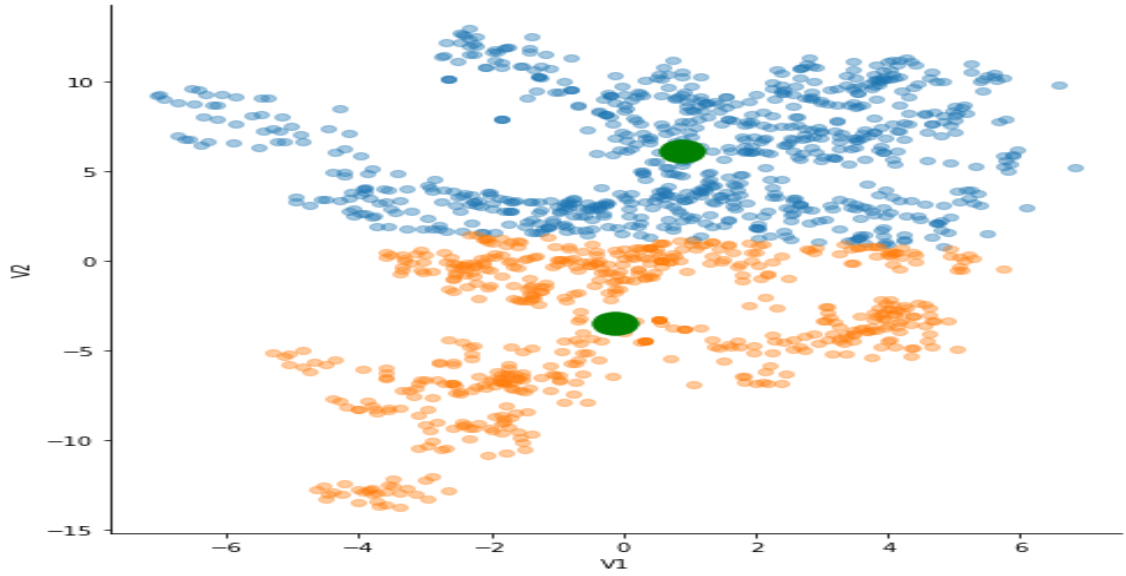
	V1	V2
count	1372.000000	1372.000000
mean	0.433735	1.922353
std	2.842763	5.869047
min	-7.042100	-13.773100
25%	-1.773000	-1.708200
50%	0.496180	2.319650
75%	2.821475	6.814625
max	6.824800	12.951600

Fig 2: The Descriptive Statistics of the Dataset

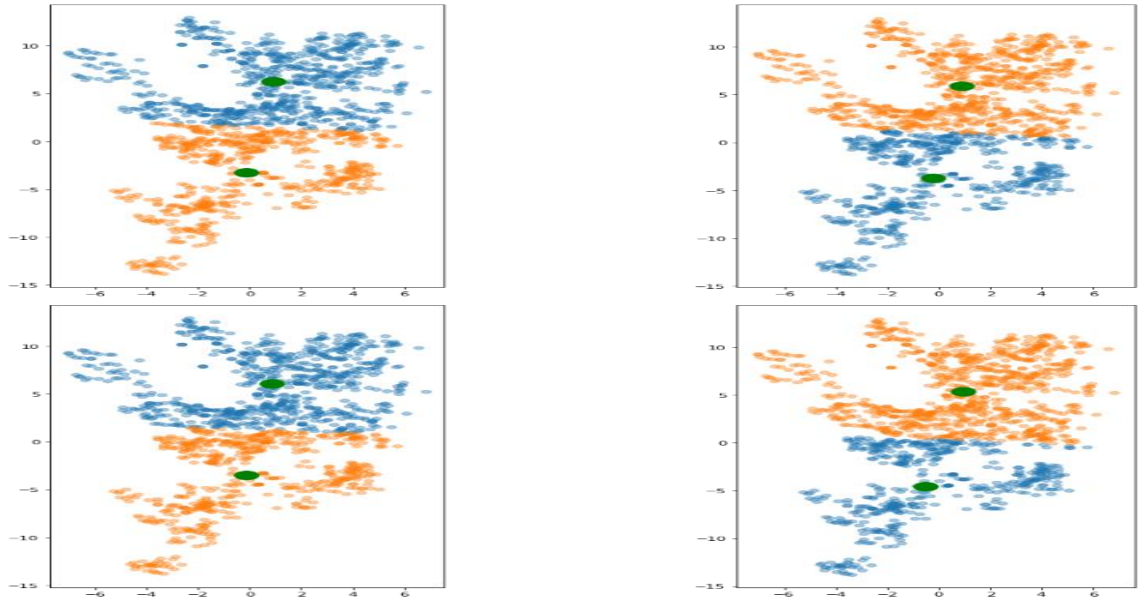
Methodology

The K-means clustering algorithm is used to find a cluster which have not been explicitly labeled in the data. A cluster is a collection of data points aggregated together because of certain similarities. The K-means algorithm start by asking the modeler the number of cluster that the modeler would like the algorithm to group. Once the modeler identifies the number of clusters, the algorithm is then allocating every data point to the nearest clusters. The center of the cluster is the mean of every data within the same cluster. Although the K-means clustering algorithm is widely used and fast to compute, it has its own limitation especially when the data contains outliers.

Modeling and Analysis: The dataset were taken into the python platform and then a popular kind of algorithm was used called K-Means Clustering algorithm which basically put N number of cluster centers (in this case 2) randomly on the data distributed freely on the graph and then using those 2 centers groups the data into 2 classes or big groups containing the values that were nearer to those points this helps us sort the data and some similar kind of values, in this project it will help us to group 2 set of values that can help us in distinguishing the forged/fake bank notes from the real ones as the data it is grouping contains the variables of the images from many note samples both fake and real that can help us in differentiating the different groups that are formed by the algorithm.



2a



2b

Fig 2a and b: Algorithm distinguishing all data values represented on the graph into 2 groups

According to the model, we have two clusters with the two centers. The center of the first cluster is approx. 0.86 and 6.04 for X-Axis and Y-Axis, respectively. The center of the second cluster is approx. -0.13 and -3.57 for X-Axis and Y-Axis, respectively. We may interpret the center of the cluster as “real” and “forged”. The dot close to the real cluster is considered as "real" by our model. The dot close to the "forged" cluster is considered as forged in our model. After running the k means algorithm several times, the clusters are quite stable. That is the estimated center of each cluster is quite close.

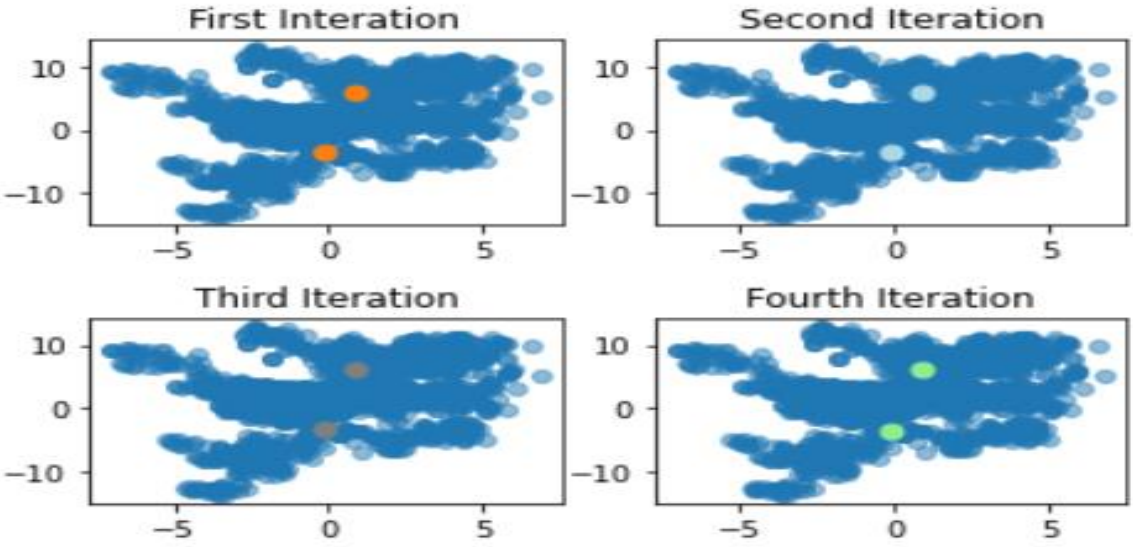


Fig 3: The scatter plot of data and the center of the two clusters in four iterations

Results

The result that we concluded was based on the multiple runs of the model that was created using the dataset, because as we know any kind of model should be stable and for the same reason it was run a total of 4 times, but it can surely be run more numbers of time to see the absolute performance of the model that was created. The result that we got was that the algorithm was able to group all the data into definite groups but in total of 4 cases the total area of grouped data in the blue color varied slightly in 2 cases and the same thing with the total area of the yellow data as seen in the graphs shown below. In the total 4 cases we ran the model we noticed similar kind of results in the 2 graphs each so, that will put us at approximately 50% efficiency, if the model was run a greater number of times the result could vary as well.

The research article demonstrates that Machine Learning can be used effectively to recognize forged banknotes. Using K-Means algorithm with several clusters equal to two can help to identify genuine/forged banknotes efficiently. Increasing the number of clusters to three can help to efficiently identify genuine/forged banknotes and a third group of banknotes which could be subject to further analysis (for example: genuine banknotes in bad conditions that can be mistaken for forged). The model has its limitation. An accuracy of 87% may be high enough in this stage. But it is still more accurate than detecting by human. At this stage, it can be used as an additional tool to reduce human error. Furthermore, by having more data and improving my analysis, the accuracy of the model can be further improved.

Conclusion

The dataset, which was given to use, we only used half of the attributes from the dataset so I will recommend that we shall use a greater number of variables gather about the image that will helps us in clearly differentiating the data.The data could be more defined and detailed as to what the values means exactly or verified range of variables could be provided about the ideal values of the variables or the attributes in the dataset, which will also lead to a better result using the algorithm.

In this research article, we propose one of the solutions to identify the forged banknotes using one of the machine learning methods called “K-Means Clustering”. According to the results, the “K-Means Clustering” seems to be a good candidate of the model to cluster the notes between “forged” and “real” as it is quite after several iterations. However, we need to check the results against the real outcomes to see if this model is suitable for this purpose.

References

1. Khan, M., et al. (2019). "Map Reduce clustering in Incremental Big Data processing." International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075 9(2): 4205-4211.

2. Priyanka B.Mohite, Prof.A.R.Kulkarni, (2016) “Enterprise Web Application Using Hadoop MapReduce System”IERJ Enterprise Web Application Using Hadoop MapReduce System” vol2 Issue 3, 1145-1149

3. Khan, M., A. Malviya and Yadav, Suryakant (2020)."BIG DATA APPROACH OF SENTIMENT ANALYSIS OF TWITTER DATA USING K-MEAN CLUSTERING APPROACH." International Journal of Mechanical and Production Engineering Research and Development (IJMPERD) 10(3): 6127-6134.

4. Khan, M. and Basheer, Shakila (2020). Using Web Log Files the Comparative Study of Big Data with Map Reduce Technique. 2020 International Conference on Intelligent Engineering and Management (ICIEM) 978-1-7281-4097-1/20/$31.00 ©2020 IEEE.

5. Khan, M., et al. (2020). Big Data and Social Media Analytics: A Challenging Approach in Processing of Big Data. Springer, Singapore. https://doi.org/10.1007/978-981-15-7961-5_59

6. Khan, M. and M. D. Ansari (2019). "Security and Privacy Issue of Big Data over the Cloud Computing: A Comprehensive Analysis." International Journal of Recent Technology and Engineering(TM) 7(6s): 413-417.

7. Aakashsoor, and Vikas, (2014),"An Improved Method for Robust and Efficient Clustering Using EM Algorithm with GaussianKernel", International Journal of Database Theory and Application vol.7, no.3, pp.191-200.

8. 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.

9. Maulik U, Bandyopadhyay S: Genetic algorithmbased clustering technique. Pattern Recognition. 2000,33: 1455-1456. 10.1016/S0031-3203(99)00137-5.

10. Alguliyev, R., Aliguliyev, R., Imamverdiyev, Y., & Sukhostat, L. (2018). Weighted Clustering for Anomaly Detection in Big Data. Statistics, Optimization & Information Computing, 6(2), 178-188. https://doi.org/10.19139/soic.v6i2.404

11. Barbara, Requirements for Clustering Data Streams, ACM SIGKDD Expl. Newsl. 3 (2003), pp. 23-27

12. G. Tzortzis and A. Likas. The MinMax K-Means Clustering Algorithm. Patt. Recog. 47 (2014), pp. 2505-2516.

13. G. Gan and M.K.-P. Ng, K-Means Clustering with Outlier Removal, Patt. Recog. Letters 90 (2017), pp. 8–14.

14. F. Jiang, G. Liu, J. Du, and Y. Sui, Initialization of K-modes Clustering using Outlier Detection Techniques. Inf. Sci. 332 (2016), pp. 167-183.

15. Shahdad, S. Y., et al. (2019). "Routing Protocols for Constraint Devices in an IEEE." IEEE, 2019. doi:10.1109/ICCSP.2019.8697933, pp. 0114-0117