Anomaly Detection Using a Fuzzy K-Means Algorithm

Avadh Naresh Kushwaha[a]

[a] *Delhi Technological University, Shahbad Daulatpur, Main Bawana, Delhi, India*

Abstract

Anomaly detection uncovers any abnormal action in the given data set for a different activity. It is an activity to finding patterns that do not conform to expected behavior. Several activities, such as the hourly temperature of a place or the number of users who visit a website per minute, have a regular pattern or waveform that depicts normal behavior. Any deviation from this normal or common pattern means that there is something unusual happening.

Several methods have been used to detect anomalies/abnormal behavior, such as clustering through finding the underlying patterns or structures within a collection of data to form different clusters. This article describes an anomaly as a particular wave shape that has never been seen before. A library of normal waveforms is created and later used to reconstruct a waveform that is being tested. The "normal" waveforms are determined using the K-means clustering algorithm to group/cluster the similar waveforms. The experiments resulted in a reconstruction error of 23.8%, which indicated there was something abnormal.

Keywords 1

K-means Algorithm, Anomaly Detection Clustering Analysis, Intrusion detection, Cluster, Outlier Detection, Waveform

1. Introduction

Anomaly detection is a method used to recognize abnormal patterns that do not match expected behavior, called outliers. Nowadays, a massive amount of data and software available on social platforms are available to ensure service. This will be required to systematically monitor the data to detect an abnormal event such as Detection of abnormal health conditions, intrusion detection, and fraud detection.

Numerous types of algorithms have been developed for anomaly detection for twin academic research, commercial interest, and the use of all this to protect such a specific application or device.

In most kinds of records such as temperature, visits on a website, etc. there is usually a normal accepted behavior indicating that everything is working fine. It might be hard to determine an anomaly based on an instantaneous value of a waveform in some cases. Instead, it's better to consider the shape of the waveform as discussed in this paper.

A library of different waveforms is created, and with the use of k-means clustering, we can develop what constitutes the normal shape of a waveform. The dataset used is plotted to form a wave of a particular shape. The waveform using the waveform library is reconstructed. If there is any poor reconstruction, then the wave is certainly containing some abnormalities.

WCNC-2021: Workshop on Computer Networks & Communications, May 01, 2021, Chennai, India.
EMAIL: avadhnareshkushwaha_2k19swe03@dtu.ac.in (Avadh Naresh Kushwaha)
ORCID: 0000-0002-2346-484X (Avadh Naresh Kushwaha)

2. Related Work

Jianliang, Haikun, and Ling [1] performed Intrusion detection on networks using K-Means. The authors' evaluated their proposed model using “the KDD Cup 1992 data”. The data provided a wide variation of instructions which were simulated over a military network. From the “100,000 labeled data items”, 1,000 of them contained attack samples, while the other 99.999 were the normal samples free from attacks.

They perform data processing for more extensive features in two ways; first, they calculate the mean absolute deviation [10] and then calculate the standardized measurement [1]. New training datasets are then created based on the formulas above, and then the proposed algorithm is applied. The experiments performed showed that the K-means algorithm was the most suitable method for intrusion detection. [1].

Xinlong and Weishi [10] used an improved k-means clustering technique to perform anomaly detection on cloud computing. The method proposed by the authors finds the known attacks and anomaly attacks in the cloud computing environment. They use a distributed intrusion model of cloud computing to perform the detections [10]. The authors also made quantitative comparisons of the different intrusion detection methods used before based on their advantages and disadvantages.

3. K-Means Algorithm

Clustering is “the process that groups objects into subclasses so that those subclasses are entirely related or have some similarities and the different clusters are altogether dissimilar from each other” [1]. The clustering algorithm is divided into four different groups, i.e., Partitioning Clustering, hierarchical Algorithm, Density-based and Grid-based.

3.1. Partitioning Clustering

Partitioning clustering [6], the data objects are directly divided into clusters of a pre-defined number. "The checking for all possible clusters is computationally impractical; certain greedy heuristics are used in the form of iterative optimization of a cluster. The partitioning clustering consists of several approaches such as K-means Clustering, K-medoid Clustering, Relocation Algorithms, and Probabilistic Clustering." [4]

3.2. Hierarchical Algorithm

A hierarchical clustering [5] "creates a hierarchical decomposition of the given set of data objects. It builds a cluster hierarchy, known as a dendrogram. Disjoint groups of clusters are obtained by cutting the tree at the desired level. Hierarchical clustering is used to find data on different levels of dissimilarity." [4]

3.3. Density-Based

Density-based algorithm [8] "can find a cluster of random shapes unexpectedly. To group objects, it uses the density objective function.in these method clusters will increase until the number of the article in the nearby growing some limitation." [4]

3.4. Grid-Based

Grid-based clustering [8] "quantizes the object space into a finite number of cells that form a grid structure. After calculating density grid-based clustering, sort the cells according to their densities. Cluster centers are identified, and all neighbor cells are traversed." [4]

Partitioning is done to divide a group of data into k-cluster [1]. K-mean constitutes a proper clustering technique by ambitious training

3.5. The Objective of Algorithm

Considering the d-dimension data set $\{x_i | x_i \in R^d, i = 1,2, \dots N\}$, the method to follow easy steps to classify given dataset, number of clusters $w_1, w_2, \dots . w_k$ to define K centroid $c = c_1, c_2, \dots . c_k$, one for each group,

$$c_i = \frac{1}{n_i} \sum_{x \in wi} x \text{ Where } n_i \text{ Are several datasets in the cluster?}$$

Select all points closest to a considering dataset and relate them to the nearest centroid. When a topic does not stand out, the first steps compete [1]. We need to the recalculated centroid of the cluster resulting from the previous actions.

A new obligate done between the current dataset and closest new centroid, an iterative loop has been generated due to this loop centroid change their location slowly until no change is done .lastly algorithm minimizing an objective function.

The objective function $J = \sum_{i=1}^{k} \sum_{j=1}^{n_i} d_{ij}(x_j, c_i)$ where $d_{ij}(x_j, c_i)$ Distance between the data point x_j and cluster center c_i.

3.6. Step by Step Algorithm

Step1: (initialize) randomly choose K occurrence $c_1, c_2, \dots . c_k$ From the data set X and start cluster center of the clustering space.

Step2: (allotment) provide each occurrence to the nearest center: $d_{ij}(X_i, C_j) < d_{ij}(X_i, C_m)$ i.e.,

$$j = 1,2 \dots . . k \ \ \& \ \ i = 1,2 \dots \dots n \ \ j \neq m \ , m = 1,2 \dots . . k \text{ and then allocate } x_i \ to \ c_j$$

Step3: (updating) Recalculate the centroid of the cluster$c_1{}^*, c_2{}^*, \dots \dots . . c_k{}^*$;

Step4: (loop) if $i = \{1 \dots . . k\}$ $c_i{}^* = c_i$ then stop the algorithm and initial $c_1{}^*, c_2{}^*, \dots \dots . . c_k{}^*$ represent the end group, or else allocate $c_i = c_i{}^*$ & perform step 2 and 3 until there isn't any more modification

3.7. Advantage(s):

K-mean algorithm is essential for a considerable large dataset [1], and its time complexity O (tkn), where t is the number of loops, k is the number of clusters, and n is the data point in the dataset

3.8. Disadvantage(s):

With the k-mean algorithm, the cluster number is required first, yet the number of clusters is usually determined later. It is sensitive to outliers [1] [2].

It is not easy to predict k-value, the k-mean algorithm with global cluster does not work well, and it does not perform well with a different group and size

4. Proposed Methodology

In our approach, we define an anomaly as a shape of the waveform that has not been experienced or seen last. The algorithm will build a normal shape library and reconstruct the waveform and determine its shape. The waveform is said to have an abnormal behavior if the reconstruction of the waveform by the library is inadequate.

4.1. Dataset

In this paper, the data used for the experiments is the EKG dataset publicly available at PhysioNet. This version of the data provides a regular waveform which provides a reasonable basis for exploring the proposed algorithm.

4.2. Clustering

To fully ascertain what constitutes a waveform's normal shape, we use k-means clustering to form different clusters to determine a normal cluster group. The clusters can be in separate n-dimensional space that's defined by n-features. In our case, the clusters are determined programmatically using the clustering algorithm.

4.3. Waveform Space

We shall consider a 32-dimensional waveform space to perform anomaly detection such that for every point within the space, it potentially represents a waveform segment. We cluster the segments that are similar to each other. "The middle of each cluster (the centroid) will measure the prototypical waveform pattern that all those segments are specific instances of". Still, where necessary, from all the waveforms, the average of them within a cluster will be considered as the centroid.

Since the center of the cluster is also a point that is part of the waveform space, this makes it also "waveform". This therefore means that, the centroids of the clusters will constitute a set of "normal waveform segments".

Now using these normal segments, we try to reconstruct the dataset being tested. The reconstruction will be considered as good if the data is similar to the original. However, suppose there is some abnormal shape in the data. In that case, it won't be possible to reconstruct the waveform using the "normal" waveform library, which then results into a reconstruction error. This therefore indicates that there is some anomaly.

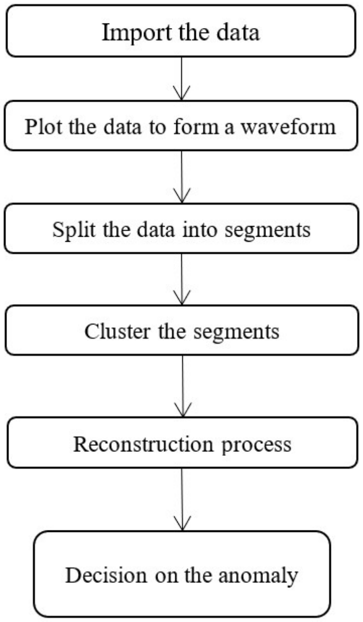
The algorithm is divided into two parts;

4.4. Training

- Divide the waveform data into segments of say n samples.
- Create a space of n dimensions with each of the segments considered as one point.
- Cluster the segment point and also determine the centroids of the clusters.
- Cluster centroids provide a library of normal waveform shapes.

4.4.1. Testing

- Using the cluster centroids in the training phase, we try to reconstruct the waveform data.
- An abnormal shape is obtained if there is any reconstruction error on any segment.



4.5. Step by Step Algorithm

Step 1: Import the data

Let FN be the filename of the data set. Let DS be the variable that holds the data after reading it using the ekg_data primitive function.

Step 2: Plot the data

Using matplotlib.pyplot function and sample size of 300, plot the data. Label the x and y axes, then show the plotted graph.

Step 3: Split data into segments

Let seg_len be the segment length of value 32, slide_len of value two, and an array named segments to store them. Loop through the DS dataset to split it into waveform segments.

Step 4: Cluster the segments

Perform clustering of the segments in a 32-dimensional space using the k-means algorithm provided in python's library scikit-learn.

The different clusters created at this point then constitutes the normal waveform shapes of the library

Step 5: Reconstruct the waveform

Try to reconstruct the data or the waveform, which will be tested using the learned library. The reconstruction process consists of four sub-steps which include;

- First, the data is split into segments that are overlapping.
- Secondly, we determine the segment's cluster centroid by using the samples' average in a cluster.
- Using the centroid from the above step, we reconstruct the segment.
- Finally, we join all the segments to form the waveform.

Step 6: Conclude on the anomaly

We decide whether there is an anomaly or not based on the reconstruction error obtained. If there is a very high reconstruction error, then we conclude that there is an anomaly.

5. Result and Analysis

Using the EKG dataset, we performed clustering on the data and also try to reconstruct to find the error between the original plotted data and the newly revamped data.

The new shape formed indicates an anomaly detected after reconstruction of the data since there is a substantial visible error. The reconstruction error obtained was 23.8%.

The figure below shows the original data's waveform, the reconstructed information, and the error after reconstruction.

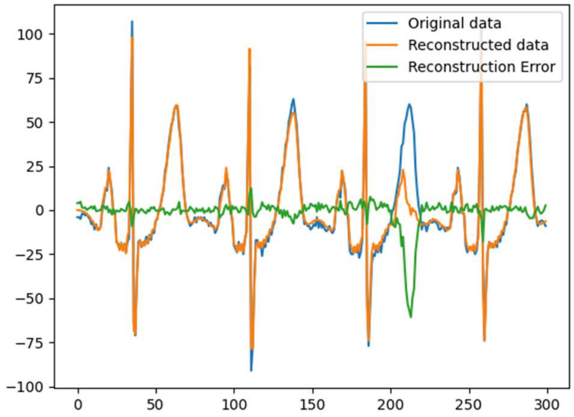


Figure 2: The waveform of original data, reconstructed data, and the error between them

6. Conclusion

Anomaly detection is essential in everyday lives, and therefore it's vital to expand on the research continuously.

Every different kind of data consists of a regular pattern that depicts normal behavior. When the data are plotted to form a waveform, it can also be hard to determine an anomaly of a waveform based on its instantaneous value but by its shape.

This paper sees how K-means clustering performs a key role in identifying normal waveform shapes in time series data to form trained library waveforms. Based on the data's reconstruction, we determine the shape of the waveform after reconstruction tested by the library created. The error obtained signifies that there is an anomaly since the shape of the waveform has not been seen before.

7. Acknowledgements

Special Appreciation to the Department of Software Engineering, Delhi Technological University, for supporting this work.

8. References

[1] M. Jianliang, S. Haikun, and B. Ling, "The Application on Intrusion Detection Based on K-means Cluster Algorithm," 2009 International Forum on Information Technology and Applications, Chengdu, China, 2009, pp. 150-152, doi: 10.1109/IFITA.2009.34

[2] M.Yazdi Pausadan, Jaka Lianto Buliali, Raden Venantivs Hari Ginardi, "Cluster Phenomenon to Determine Anomaly detection on Flight Riute," the fifth Information System International Conference 2019.

[3] C. Yin, S. Zhang, J. Wang, and J. Kim, "An Improved K-Means Using in Anomaly Detection," 2015 First International Conference on Computational Intelligence Theory, Systems and Applications (CCITSA), Ilan, Taiwan, 2015, pp. 129-132, doi: 10.1109/CCITSA.2015.11.

[4] C. T. Baviskar and S. S. Patil, "Improvement of data object's membership by using Fuzzy K-Means clustering approach," 2016 International Conference on Computation of Power, Energy Information and Communication (ICCPEIC), Melmaruvathur, India, 2016, pp. 139-147, doi: 10.1109/ICCPEIC.2016.7557238.

[5] I.P and G.DK, "A Survey on different clustering algorithm in data mining technique" International Journal of Modern Engineering Research (IJMER), Vol.3, no.1, pp.267-274, 2013.

[6] S. Ayramo and T. Karkkainen, "Introduction to partitioning based clustering methods with a robust example," Reports of the Department of Mathematical Information Technology Series C. Software and Computational Engineering, 2006.

[7] E. G. Mansoori, "Frbc: a fuzzy rule-based clustering algorithm," Fuzzy Systems, IEEE Transactions on, vol. 19, no. 5, pp. 960-971, 2011.

[8] J. Han and M. Kamber, "Data mining: concepts and techniques," United States of America: Morgan Kauff Mann Publishers, 2001.

[9] C. C. Aggarwal and C. K. Reddy, Eds., Data Clustering: Algorithms and Applications. CRC Press, 2014. [Online]. Available: http://www. Charu Aggarwal. netlclusterbook.pdf.

[10] X. Zhao and W. Zhang, “An Anomay Intrusion Detection Method Based on Improved K-Means of Cloud Computing” 2016.