ABSTRACT

This paper provides an overview of anomaly and various anomaly detection methods. Since the data of this paper is unlabeled, the focus is on unsupervised anomaly detection methods with various literature reviews of previous studies that outlies pros and cons of each method. Furthermore, it proposes a novel algorithm, demonstrated by pseudocode, to detect anomalies based on the K-mean and statistical methods. In proposed algorithm, the parametric statistical method with the threshold of anomaly score is applied on the univariate data of K-means minimum distance to determine anomalous points. Then, the most important features are selected through PCA and the probability of best feature selections and then visualize anomalies on these most important feature selections.

Key Words: Anomaly Detection, Clustering, K-means, PCA, Probability of Feature Selection, Anomaly Score, Algorithm and Pseudocode, Visualization, Python and R Programming

1-What is anomaly detection

In simple words, anomaly detection is a process of identifying rare observations in a dataset that do not follow the main pattern in its local. As **Error! Reference source not found.** shows, F1, F2 and F3 deviate from most data and do not follow the pattern in G1 and G2.

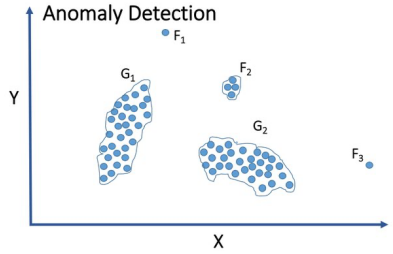

Figure 1- Anomaly Detection. (Domingues, 2022)

Anomaly detection can address many problems in the world of fraud, finance, medicine, etc. In the medical field, anomaly detection can be applied to identify any time periods of unusual beats. Recently Electrocardiography (ECG) anomaly detection has become a prevalent task among researchers. Pavel Senin, et al (2015) examined anomaly detection to identify disease ECG signals to measure the health of the human heart. They propose a novel method for obtaining ECG time series data and identifying the anomaly heart signal by comparing the similarity. This ECG visualization is illustrated in Figure 2 (Senin, et al., 2015).

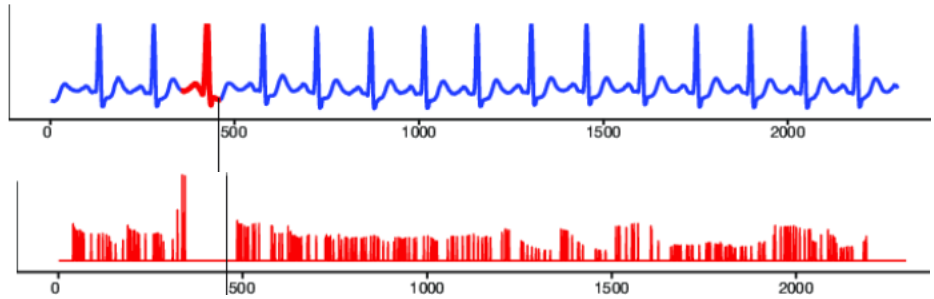

Figure 2- Anomaly Detection in ECG Time Series Data (Senin, et al., 2015)

Methods of Anomaly Detection

There are different methods for anomaly detections that are in the Figure 3:

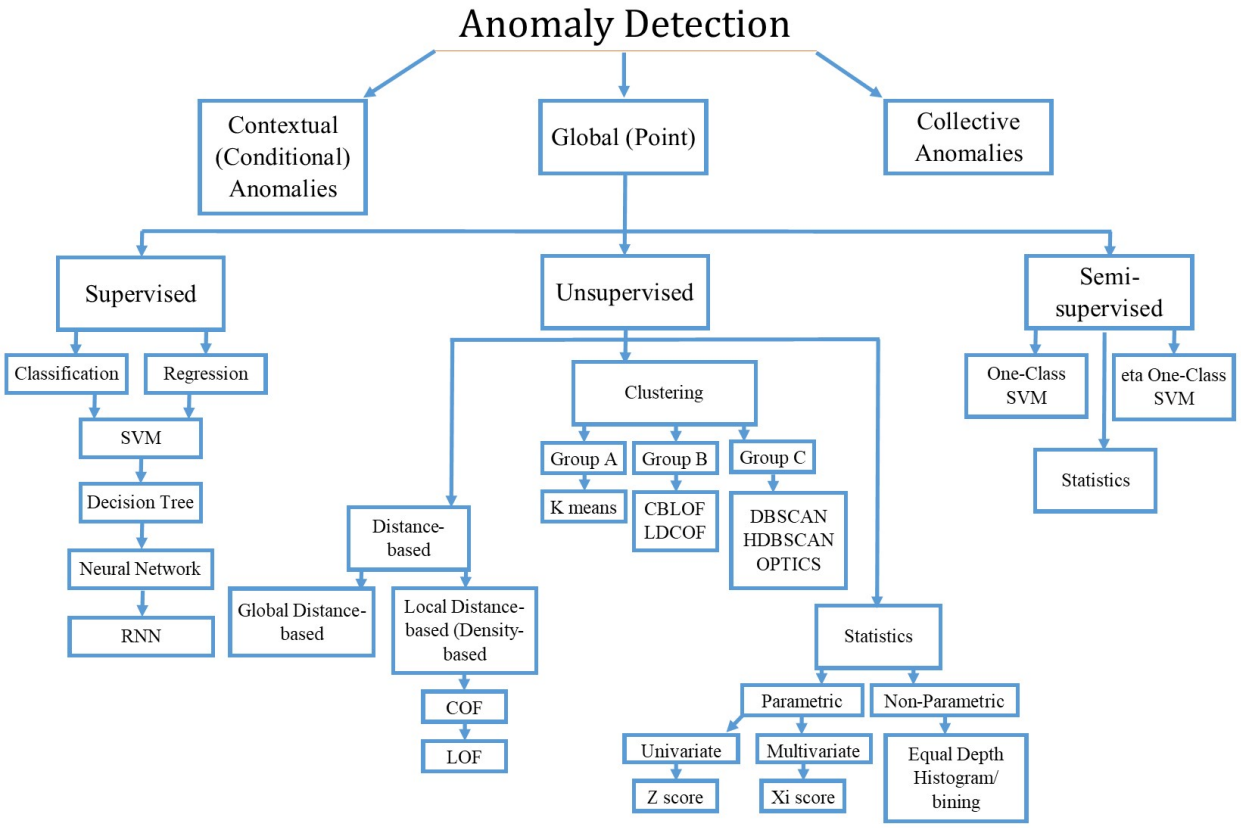

Figure 3-Anomaly Detection Methods in Supervised, Unsupervised and Semi-Supervised Algorithm.

Since the data is unlabeled, the unsupervised methods of the Figure 3 are emphasized in this paper as follows:

1) Distance-Based Methods:

It measures the relative distance to the K nearest neighbors that is defined by radius. This measures how close the datapoints are to their neighbors. According to Golstein and Uchida (2016), the distance-based method is classified to global distance-based and local distance-based outlier detection.

1-1) Global Distance-Based Outlier Detection:

KNN is the most popular Euclidean method of global distance-based outlier detection.

1-1-1) KNN Outlier Detection: According to Gu et all (2019), In this method, KNN computes the average distance to K nearest neighbors for each data point. This is ratio of Dk (xi) to average Dk (xj) for its neighbours, illustrated in (Equation 1).

$$D_k(x_i) = \frac{1}{k}\sum_{j\in N_k(x_i)}\|x_i - x_j\|$$

(Equation 1)

In the above equation, $N_k(x_i)$ be the k-nearest neighbours of x_i. $D_k(x_i)$ is the average distance to k-nearest neighbours.

$$A_k(x_i) = \frac{D_k(x_i)}{\frac{1}{k}\sum_{j \in N_k(x_i)} D_k(x_j)}$$

(Equation 2)

(Gu, Akoglu, & Rinaldo, 2019)

If $A_k(x_i)$, that is, the result of (Equation 2) is greater than 1, x_i is outlier as it is further away from its neighbours. These x_i points that are global outliers are illustrated with red colors in Figure 4.

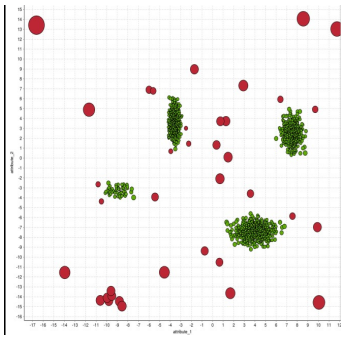

Figure 4 - Visualization of K-NN Global Anomaly Detection Algorithm. (Goldstein & Uchida, 2016)

1-2) Local Distance-Based Outlier Detection (Density-Based methods):

The local density-based outlier detection is illustrated in Figure 5.

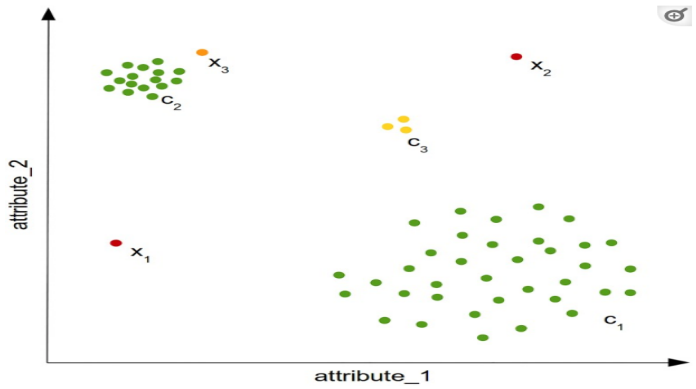

Figure 5- Visualization of Global Anomalies as (x1,x2), Local Anomaly x3 and Micro-Cluster c3. (Goldstein & Uchida, 2016)

Outlier x3 has similar density as elements of cluster c1. The solution for this local density-based outlier detection is LOF.

1-2-1) Local Outlier Facto (LOF): It compares the density of any given point to the density of its neighbors and determine if the data point is considered anomaly or normal. As a rule of thumb, the normal data points have values between 1 and 1.5, while anomalous points are higher. Alghushairy et all

(2021) provide (Equation 3), (Equation 4), **Error! Reference source not found.**, **Error! Reference source not found.** and Figure 6 as follows

The local reachability density (Lrd) of data point p is defined in (Equation 3)

$$Lrd_{MinPts}(p) = 1/\left(\frac{\sum_{o\in N_{MinPts}(p)} reach-dist_{MinPts}(p,o)}{|MinPts(p)|}\right)$$

(Equation 3)

In Equation, the average reachability distance based on Minpts number of nearest neighbor of data point p is calculated.

From the above equation, the LOF score of data point p is calculated as illustrated in (Equation 4). This equation calculates the average ratio of the local reachability density of data point p and the Minpts-nearest neighbor of data point p.

$$LOF_{MinPts}(p) = \frac{\sum_{n\in N_{MinPts}(p)} \frac{Lrd_{MinPts}(o)}{Lrd_{MinPts}(p)}}{|N_{MinPts}(p)|}$$

(Equation 4)

To determine whether data point p is threshold or not, the threshold score θ is used.

(Alghushairy, Alsini, Soule, & Ma, 2021)

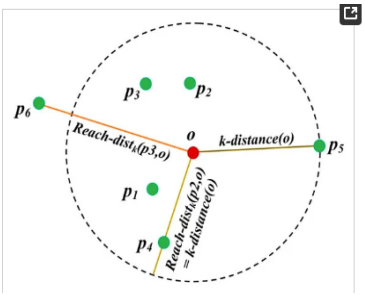

Figure 6. The Reachability Distance for Various Data Points of p with Regard to Center o. (Alghushairy, Alsini, Soule, & Ma, 2021)

2-Statistical methods

Statistical methods are categorized into two groups that are Parametric and non-parametric methods.

2-1) Parametric methods: Parametric methods examine the parametric distribution, and the probability of any object in the distribution implies the anomaly score. Parametric methods can be applied to univariate and multivariate data.

2-1-1) Parametric methods to Univariate data:

Univariate data involves only one attribute or variable. Parametric method to univariate distribution is calculated by normal distribution (i.e., Gaussian distribution) and Z score, illustrated in (Equation 5) (Equation 6, (Equation 7, (Equation 8 and Figure 7. In (Equation 6 and (Equation 7, μ and σ are mean and standard deviation, respectively.

$$N(\mu,\ \sigma^2)\colon f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-u}{\sigma}\right)^2}$$

(Equation 5)

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

(Equation 6)

$$\boldsymbol{\sigma} = \sqrt{\frac{\sum(\boldsymbol{x_i} - \boldsymbol{\mu})^2}{\boldsymbol{N}}}$$

(Equation 7)

$$Z = \frac{(x_i - \mu)}{\sigma}$$

(Equation 8)

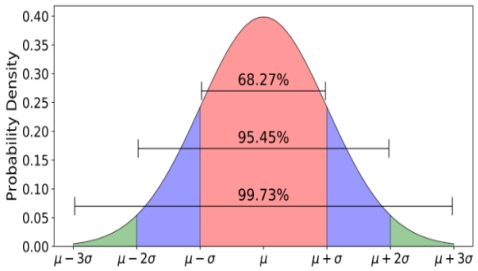

Figure 7: Normal Distribution (Gaussian distribution) with Regard to μ and σ. (McLeod, 2019)

2-1-1-1) Anomaly score: It defines the likelihood that an object is an anomaly. So, based on this definition, Z score in Figure 7 of univariate distribution is anomaly score.
Figure 7 indicates a proven statistic rule. The probability that $x \sim N(\mu, \sigma)$ fall between $\mu - 3\sigma$ *and* $\mu + 3\sigma$ is almost 0.9973. This implies the Z *score or anomaly score* $\in [-3,3]$. As a result, when data points fall out of this range of anomaly score $[-3,3]$, these points are considered as anomalies.

2-1-2) parametric methods to multivariate data:

Multivariate data involves two or more attributes or variables. Parametric method to multivariate data is calculated by (Equation 9.

$$x^2 = \sum_{i=1}^{n} \frac{(o_i - E_i)^2}{E_i}$$

(Equation 9)

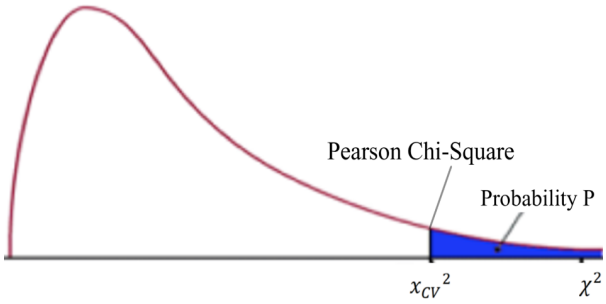

Figure 8- Chi-Squared Distribution of Multivariate Data, (Saylor Academy, 2012)

As Equation 10, E_i is the mean of the *i*-th dimension among all objects (i.e., expected value), o_i is value of *o* in the *i*-th dimension (i.e., observation), and *n* is the dimensionality. (Equation 9 is considered anomaly score in multivariate data. As Figure 8, χ^2 statistic (anomaly score) is larger than Pearson Chi-Square statistic (χ_{cv}^2), therefore thw probability p of observations o_i are considered as anomalies.

2-2) Non-parametric methods:
For non-parametric methods, a counting based histogram is considered to detect outliers as it is shown in Figure 9. Anomalies are found through equal-depth histogram/binning that indicate finding outliers through smoothing by bin mean, bin median and bin boundary. Split the number of records for bins of the histogram and apply some threshold to filter anomalies.

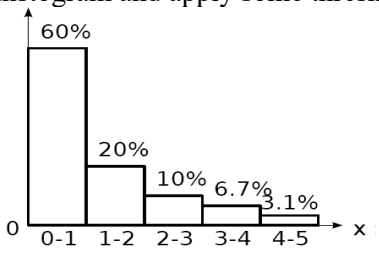

Figure 9: Non-parametric Histogram Distribution

3) Clustering-Based methods

There are Two clustering-based methods that include outliers do not belong to clusters (i.e., Group A and Group B) and outliers belong to clusters (i.e., Group C).

3-1) Group A:

Outliers belong to the cluster and considers the distance between the object and its cluster center to detect anomaly. Hussain et al. (2020) described the outlier removal clustering (ORC) to detect outliers. They explain that ORC needs two steps to detect anomalies, including 1)- K-means clustering and 2)- Detect the data points far away from cluster centroids. This is demonstrated in (Equation 10 and (Equation 11.

$$d_{max} = max_x\{\|x_i - c_{pi}\|\}$$

(Equation 10)

$$o_i = \frac{\|x_i - c_{pi}\|}{d_{max}}$$

(Equation 11)
(Hussain, 2020)

In the above equations, x_i is the i-th observation (i = 1,2,3…. N), dmax is the maximum distance of observation from its centroid (Cpi) and oi is outlier score with a value between 0 and 1. If the result of (Equation 11 is greater than a specific threshold, then observations o_i are considered anomalies.

3-1-1) Pros and Cons: The advantage of this method is it is distance-based approach that uses Euclidean distance and removes the potential outlier clusters based on the distance threshold. The weakness is that it does not consider the densities and the potential of small clusters to be outliers.

3-2) Group B:

Outliers belong to the cluster and examine a small set of objects close to the large cluster as an anomaly. According to Amer and Goldstein (2012), Cluster-Based Local Outlier Factor (CBLOF) algorithm estimates the region of similar densities by using the concepts of local density from K-nearest neighbors.

CBLOF compares the densities of data points to their neighbourhood by calculating the deviation of each data point to the cluster centroid. The CBLOF score for data instance (p) is defined in(Equation 12. This equation indicates that CBLOF of data point p, for instance, is calculated by multiplying the distance of the data point (p) to the center of its nearest large cluster by the number of data points that belong to the cluster. If the data point is in a small set, the distance to the relatively large cluster is used, while if the data point is in a larger cluster, the distance to the center is applied.
Figure 10 illustrates that point (p) is in the small cluster C2, and thus the outlier score will be equal to the distance to the large cluster. The nearest large group is C1 which is multiplied by 5, which is the size of C2.

$$CBLOF(p) = \begin{cases} |C_i| \cdot \min(d(p, C_j)) \text{ if } C_i \in SC \text{ where } p \in C_i \text{ and } C_j \in LC \\ |C_i| \cdot d(p, C_i) \text{ if } C_i \in LC \text{ where } p \in C_i \end{cases}$$

(Equation 12)
(Amer & Goldstein, 2012)

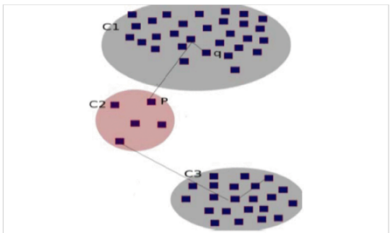

Figure 10- Cluster-Based Local Outlier Factor. (Amer & Goldstein, 2012)

3-2-1) Pros and Cons: The weakness of CBLOF is it works at a global level and uses the number of clustering members and does not consider the density of the cluster and spherical distributions.

3-2-2) Local Density Cluster-based Outlier Factor (LDCOF): LDCOF resolves the CBLOF issue by considering an estimation of the cluster densities and a spherical distribution for the cluster members. According to Amer and Goldstein (2012) and Katz (2020), LDCOF follows a similar approach to CBLOF besides the normalization of outlier score. (Equation 13 demonstrated the distance average of cluster C. In this equation, d(i, C) shows the distance between a point p and the center of cluster C.

$$distance_{avg}(C) = \frac{\sum_{i \in C} d(i, C)}{|C|}$$

(Equation 13)
(Katz, 2020)
(Equation 14 denotes the normalized LDCOF score that is the distance to the nearest large cluster divided by the distance average in the cluster C.

$$LDCOF(p) = \begin{cases} \frac{\min(d(p,C_j))}{distance_{avg}(C_j)} \text{ if } p \in C_i \in SC \text{ where } C_j \in LC \\ \frac{d(p,C_i)}{distance_{avg}(C_i)} \text{ if } p \in C_i \in LC \end{cases}$$

(Equation 14)
(Amer & Goldstein, 2012)

The higher score of CBLOF and LDCOF indicate that the data points are further away and are more anomalous.

3-3) Group C:

Outliers do not belong to any cluster. DBSCAN algorithm is used to distinguish the anomalous cluster from normal cluster and increase the accuracy of clusters. According to Nanehkaran, et al (2022), DBSCAN has two parameters that are minimum number of points (MinPts) and neighborhood radius (Eps). The DBSCAN clustering algorithm is applied and parameters (Eps) and MinPts are optimized.

(Equation 15(Equation 16 demonstrate the neighborhood of desired p and radius is specified. (Equation 16 indicates the neighborhood of desired p must be larger than MinPts. In these equations, D is all data points. If these conditions in (Equation 15(Equation 16 are met, the new normal cluster is created and outliers are distinguished from the new clusters. Figure 11 shows that how parameters of (Eps) and MinPts contribute to each other to define clusters and outliers. In this figure, the distribution of ϵ is based on the maximum Euclidean distance between feature vectors.

$$N_{Eps} = \{q \in \frac{D}{dis(p,q)} < Eps\}$$

(Equation 15)

$N_{Eps}(p)$ > MinPts

(Equation 16)

(Nanehkaran, et al., 2022)

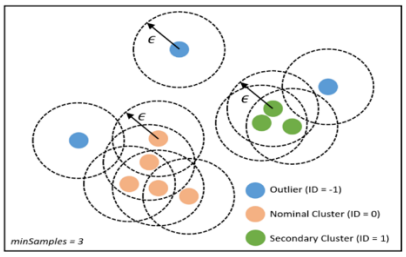

Figure 11- DBSCAN Clustering Visualization (Sheridan, et al., 2020)

4) Methodology and Designing Algorithm

4-1) Using K-means Algorithm to Detect Anomalies:

The data points of this study (wine data) have 11 features and has spherical shape. According to Elbow method and Silhouette method in Figure 12 and **Error! Reference source not found.**Figure 13, the data set has 3 K-means cluster. Both methods confirm the number of clusters are 3.
The histogram of these clusters is illustrated in Figure 14. To find anomalous points, we use the group A of clustering that is combination of K-means and parametric threshold.

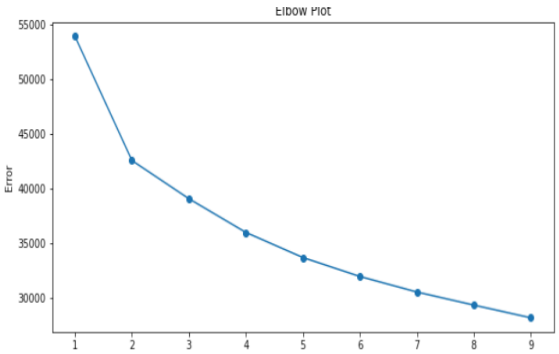

Figure 12- Sum of Square Elbow Method to Determine Number of Clusters.

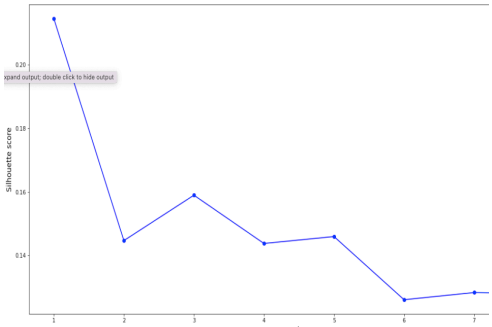

Figure 13 - Silhouette Coefficients to Confirm the Number of Clusters.

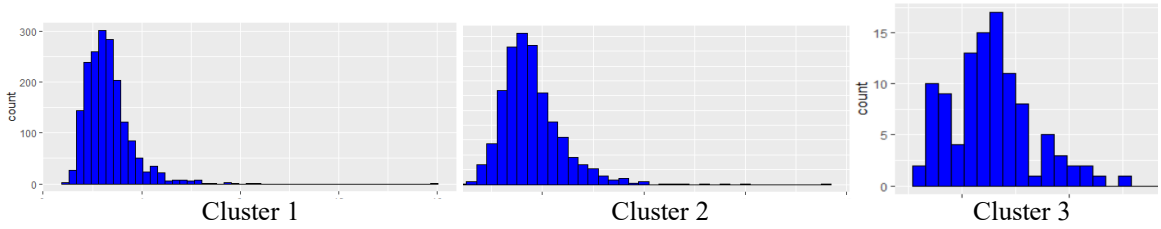

Figure 14- Histogram Visualizations of Cluster 1, Cluster 2 and Cluster 3, determined by **Error! Reference source not found.** and Figure 13.

4-2) Pseudocode of Designing Algorithm:

As for designing this algorithm, it is required to provide the assumption and threshold. According to Virmani et all (2015), K-means algorithm is probable to irregularities with different cluster sizes. Therefore, the assumption (i.e., null hypothesis) of this algorithm is all clusters have the equal size. The assumption of threshold is anomaly score. This anomaly score is obtained by the parametric statistics on univariate datasets that is Z score.

Assumption:

size(C1) = size(C2) =…= size (Ck)

Input:

1) D: A data set include both cluster labels and distances calculated with k-means algorithm
2) Threshold_Z_Score = Threshold to measure the variability of each data point.
3) Average_cluster_size = divide the count of all data points by count of clusters

4) Threshold_for_cluster = A ratio of size of cluster k to average size of clusters (Average_cluster_size)

Output: Anomalies

begin

```
5)  for cluster k in all clusters do
6)      if |k| / Average_cluster_size < Threshold_for_cluster # |k| is data points in cluster k
7)          consider all data points in k as an anomaly
8)      else
9)          for each point x in cluster k do
10)             Calculate Z score of x
11)             if z-score(x) > Threshold_Z_Score
12)                 consider x as an anomaly
13)         end for
14) end for
```

Line 1: Set the input of anomaly detection algorithm. This input is K-means distances and cluster labels. Cluster labels vary from number 1 to number k (k=1,2, 3,…, K) and distance is a distance of each data point to center of cluster k.

Line 2: Set the Threshold_Z_Score manually that is anomaly score. Although the Z-score range is [-3,3], Threshold_Z_Score is set manually because the number of data points can impact this threshold.

Line 3, 4: Set the Average_cluster_size and Threshold_for_cluster.

Line 5, 6, 7: Filtered the anomalous clusters that their normalized size is less than Threshold_for_cluster. This is with the assumption of equal cluster size.

Line 10: Calculated the Z score for each data point.

Line 11, 12: Filtered the anomalous data points that their Z score is greater than Threshold_Z_Score.

4-3) Implementation of the proposed anomaly detection model and Visualize Anomalies

After applying the above algorithm that is the combination of K-means and threshold on the data sets, the anomalous points are detected. To visualize the anomalies on the data set, it is required to find the most important features of data. The most important features are selected based on two methods that are advanced PCA feature selection and selectkbest algorithim that determine the probability score for features. Both methods confirmed the same features as the most important features in the dataset that are illustrated in Table 1.

	PC1	PC2	PC3	PC4	PC5	PC6
density	0.473304	0.057188	-0.112468	0.004673	-0.089899	0.339204
residual sugar	0.401503	0.042162	-0.205346	-0.252174	0.004664	0.305489

Specs	Score
density	165.534748
residual sugar	163.629560

Table 1. Determining the Most Important Features Through PCA and selectkbest algorithm.

4-4) Visualization of Anomalous Points

The visualization of anomalous points based on the proposed algorithm on the most important features is illustrated in Figure 15.

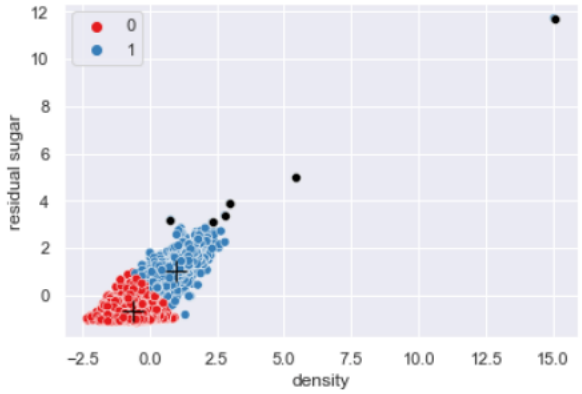

Figure 15- Visualization of the Anomalous Points on the Most Important Features

Conclusion

Since the data has a spherical shape, the K-means is used to cluster the data. However, K-means is unable to detect anomalies and noise. Therefore, the proposed algorithm in (4-3) Implementation of the proposed anomaly detection model and Visualize Anomalies is used to filter the data points and clusters. This proposed algorithm uses Group A of clustering in (3-1) **Group A**: that utilizes the combination of statistical methods and K-means clustering algorithm to determine anomalies. The statistical method of Group A is applying the parametric method to univariate data points because K-means minimum distance to each cluster centroid is univariate. The threshold for cluster size of proposed algorithm is applying cluster threshold for the K-means normalized cluster (ratio of the size of the K-means cluster to the average size of clusters) to detect the anomalous clusters. The K-means clustering was performed in Python, where we find the most important features based on the PCA and selectkbest algorithm. Then, the anomalous points obtained by the proposed algorithm visualized on the most important features of data.

Bibliography

Alghushairy, O., Alsini, R., Soule, T., & Ma, X. (2021). A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams. *Big Data and Cognitive Computing*, 5(1).

Amer, M., & Goldstein, M. (2012). Nearest-neighbor and clustering based anomaly detection algorithms for rapidminer. *Proc. of the 3rd RapidMiner Community Meeting and Conference* (pp. 1-12). RCOMM.

Domingues, R. (2022, June 20). *Machine Learning for Unsupervised Fraud Detection.* Retrieved from Digitala Vetenskapliga Arkivet: https://www.diva-portal.org/smash/get/diva2:897808/FULLTEXT01.pdf

Gu, X., Akoglu, L., & Rinaldo, A. (2019). Statistical Analysis of Nearest Neighbor Methods for Anomaly Detection. *33rd Conference on Neural Information Processing Systems.* Vancouver, Canada: NeurIPS .

Hussain, S. M. (2020). A novel unsupervised feature-based approach for electricity theft detection using robust PCA and outlier removal clustering algorithm. *International Transactions on Electrical Energy Systems, vol. 30, no. 11*, 14-19.

Katz, D. (2020). *Identification of Software Failures in Complex Systems Using Low-Level Execution Data.* Retrieved from Deby Katz: http://www.debykatz.com/dsk_thesis.pdf

McLeod, S. (2019). *Z-Score: Definition, Calculation and Interpretation*. Retrieved from SimplyPsychology: https://www.simplypsychology.org/z-score.html

Nanehkaran, Y. A., Licai, Z., Chen, J., Jamel, A., Shengnan, Z., Dorostkar Navaei, Y., & Abdollahzadeh Aghbolagh, M. (2022). Anomaly Detection in Heart Disease Using a Density-Based Unsupervised Approach. *Wireless Communications and Mobile Computing*.

Saylor Academy. (2012). *Chi-Square Tests for Independence*. Retrieved from Introductory Statistics : https://saylordotorg.github.io/text_introductory-statistics/s15-01-chi-square-tests-for-independe.html

Senin, P., Lin, J., Wang, X., Oates, T., Gandhi, S., Boedihardjo, A., . . . Frankenstein, S. (2015). Time series anomaly discovery with grammar-based compression. *18th International Conference on Extending Database Technology (EDBT),* (pp. 481-492). Brussels, Belgium: OpenProceedings.org.

Sheridan, K., Puranik, T. G., Mangortey, E., Pinon, O. J., Kirby, M., & Mavris, D. N. (2020). An application of dbscan clustering for flight anomaly detection during the approach phase. *AIAA Scitech 2020 Forum* (p. 1851). Atlanta, GA: Georgia Institute of Technology.

Virmani, D., Taneja, S., & Malhotra, G. (2015). Normalization based K means Clustering Algorithm. *International Journal of Advanced Engineering Research and Science*, 1-5.