Anomaly Detection of Boarding Patterns

Burak Keskin
Computer Engineering Department
Akdeniz University
Antalya, Turkey
20195156008@ogr.akdeniz.edu.tr

***Abstract*—In this paper, a review on clustering algorithms, anomaly detection and analysis of the given data is presented. Literature of clustering algorithms and anomaly detection is classified according to the basis of the algorithms. In each classified section, methods used in the classified algorithm has been explained. It is observed that usage and clustering algorithms in anomaly detection techniques is practical for finding accurate outliers in the given data.**

***Index Terms*—Data clustering, outlier, density-based clustering, hierarchical clustering, K-Modes Clustering, DBscan Clustering**

I. Introduction

Anomaly detection is an important concept to find interesting information about the data. Most of the anomaly detection techniques are developed for certain application domains, however there are also techniques for general usage areas. Anomaly detection techniques are used to find the exceptions, outliers, surprises in the data. In order to find outliers, according to the type of the data in the dataset, different clustering techniques can be applied.

To determine which clustering technique is going to be used, data should be analyzed first and for example if the dataset contains only numeric values, K-Means algorithm is suitable for detecting anomalies. However, modern day datasets are mostly combined of numeric and categorical values so techniques that applied to numerical and categorical data should be combined into new techniques in order to analyzed mixed type datasets. K-Modes technique, proposed by Huang in 1998, modifies K-Means clustering technique to be able to applied on the categorical data.

In this paper, clustering techniques that can be used to detect anomalies are explained in the following sections. In this assignment, dataset is a mixed type dataset hence the focus will be understanding the K-Modes and KPrototypes techniques,

II. Objective

In this assignment, objective is to analyze the given dataset of passenger boarding information on 18/12/2018 and find the anomalies on this dataset. Initial dataset consists of three columns: Line, BoardingTime and PassengerCount. PassengerCount is the numerical value and Line,BoardingTime are the categorical values. That means we have mixed type dataset. We have 51 unique bus lines in our dataset. In order to analyze the dataset, first data is needed to be organized. Screenshot of the sample of initial dataset is shown in Figure 1.

	A	B	C	D
1	Line	BoardingTi	PassengerCount	
2	KC06	2019-12-1	1	
3	LF09	2019-12-1	3	
4	TC93	2019-12-1	1	
5	TL94	2019-12-1	1	
6	TL94	2019-12-1	1	
7	VF01	2019-12-1	4	
8	AC03	2019-12-1	7	
9	AF04	2019-12-1	1	
10	DC15	2019-12-1	33	
11	DC15A	2019-12-1	10	
12	FL82	2019-12-1	6	
13	KC06	2019-12-1	3	
14	KL08	2019-12-1	21	
15	LF09	2019-12-1	4	
16	MF40	2019-12-1	1	
17	TC93	2019-12-1	6	
18	TCP45	2019-12-1	1	
19	VF01	2019-12-1	20	
20	VF02	2019-12-1	4	

Fig. 1. Initial Dataset

Column BoardingTime consists of both the day and time data. It is better to separate those two data to use both features in our analysis. Using python's pandas library, BoardingTime column is separated from the T letter and added to the dataset as two new columns named Day, Time. Screenshot of the sample of final dataset that is going to be used in the analysis is shown in Figure 2. Now the dataset is ready to be processed but before processing the dataset, literature review is needed to determine which techniques can be applied to the dataset and what are the advantages and disadvantages of each technique to obtain the knowledge and use that knowledge in further studies too. In the literature review, clustering techniques are reviewed first and anomaly detection detection techniques are reviewed after clustering techniques. Even if clustering techniques that are used for only numerical data is not suitable for the dataset used in this assignment, most popular among them are also reviewed in order to observe why they are not suitable. As stated in the introduction section, technique that is going to be used for analyzing the dataset is going to be K-Modes and KProtoypes techniques. For comparison, KMeans, DBScan and BIRCH methods are also going to be explained.

dfnew - DataFrame

Index	Line	sengerCo	Day	Time
0	KC06	1	20191218	0000
1	LF09	3	20191218	0000
2	TC93	1	20191218	0000
3	TL94	1	20191218	0000
4	TL94	1	20191218	0100
5	VF01	4	20191218	0330
6	AC03	7	20191218	0530
7	AF04	1	20191218	0530
8	DC15	33	20191218	0530
9	DC15A	10	20191218	0530
10	FL82	6	20191218	0530
11	KC06	3	20191218	0530

Fig. 2. Final Dataset

III. Literature Review

Clustering is the division of the data into similar object groups. Each cluster consists of similar objects but different clusters contain dissimilar objects. Fewer clusters provides simplicity but may cause loss in fine details about the data.

According to the paper [1], there are three main classification categories for clustering techniques. These techniques are partitional clustering, hierarchical clustering and density-based clustering. However, these main classification categories have been expanded by a newer paper [2]. According to the [2], there are two more classification of clustering techniques which are grid-based and model-based. In this paper, partitional, hierarchical and density-based clustering techniques will be discussed. In figure 3 , classification schema of clustering techniques according to the [2] are shown.

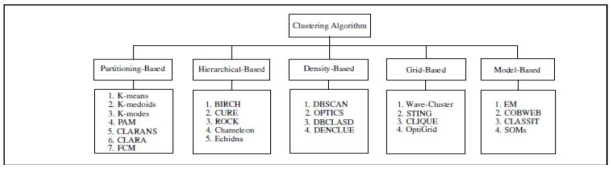

Fig. 3. Classification of Clustering Algorithms

A. Partition-Based Clustering

Partition-based algorithms are divided into two main categories, the centroid and medoid algorithms. Centroid algorithms represent the clusters by using their gravity center. Medoid algorithms represent the clusters by the means of the closest to the gravity center.Most commonly used partition-based clustering algorithm is k-means algorithm.

1) K-Means: K-means partitions the dataset into k number of subsets which all the the subsets,clusters are closest to the same centre [1]. K-means algorithm's working principle is summarized according to the paper [3] below:

- Step 1: Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
- Step 2: Assign each object to the group that has the closest centroid.
- Step 3: When all objects have been assigned, recalculate the positions of the K centroids.
- Step 4: Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

K-means method is suitable for numerical data because it takes the mean of the data points and partitions clusters according to the mean. However, when it comes to categorical data, since there is no numerical value in categorical data, k-means method is not applicable. Instead of k-means, methods based on k-medoids are applicable to the categorical data.

2) K-Modes: K-modes algorithm is variant of k-means algorithm that eliminates the limitations of k-means [4]. In k-means, means of clusters were used to distinguish the clusters, in k-modes, as its name on it, mode of the clusters are used as dissimilarity measurement. It is suitable for clustering categorical data, but in rare occasions k-modes may suffer from accuracy [5].

B. Hierarchical-Based Clustering

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Hierarchical clustering begins with treating each data point as separate clusters. After each data point is marked as separate cluster, combine the similar clusters together until there is no cluster left to combine. Hierarchical clustering is divided into two main categories, agglomerative and divisive.

1) Agglomerative: In agglomerative clustering, final result is one big cluster of data. BIRCH is an example of agglomerative clustering technique. Steps of agglomerative clustering is stated below:

- Calculate the similarity of one cluster with all the other clusters (calculate proximity matrix).
- Step 1: Consider every data point as a individual cluster.
- Step 2: Merge the clusters which are highly similar or close to each other.
- Step 3: Recalculate the proximity matrix for each cluster.
- Step 4: Repeat Step 3 and 4 until only a single cluster remains.

a) BIRCH: BIRCH is an agglomerative clustering algorithm that is commonly used. It is claimed to be the first clustering algorithm that can handle noisy data [4]. Algorithm is explained below:

- Data is scanned and CF-tree(Clustering Feature Tree) is built.
- Global or semi-global clustering algorithms are applied to the CF-tree.
- If accuracy is not at the desired level, repeat steps 1 and 2 to increase accuracy.

2) Divisive: In divisive clustering, algorithm is the opposite of agglomerative clustering.

- Complete dataset is taken as cluster at first.

- In every iteration, dataset is divided into smaller clusters.
- At the end, there are N number of clusters ready to be analyzed.

C. Density-Based Clustering

Density-based clustering is the technique that separates the clusters according to their densities. Meaning specific cluster is surrounded by other clusters with lower density than the specific cluster. If a point in the dataset is not in the limits of threshold, it can not be a member of the cluster thus considered as an anomaly. DBSCAN is one of the most commonly used method in density-based clustering.

1) DBSCAN: Algorithm of DBSCAN is explained below [4]:

- Step 1: Select an arbitrary point p.
- Step 2: Find all directly reachable points from point p.
- Step 3: If p is border point, no point will be able to reach p. Hence select next point in dataset as p.
- Step 4: If p is core point, that means it is a cluster.
- Step 5: Continue above steps until all points in the dataset are covered.

a) Advantages of DBSCAN:

- It does not need to define number of clusters like K-means.
- Clusters are good quality independent of their shape.
- It is robust to outliers.

b) Disadvantages of DBSCAN:

- It has two input parameters: eps value and arbitrary point p. Quality of clusters are depending on these two variables.
- It works only on numerical dataset.
- Identifying adjacent clusters with different density is a problem.

D. Anomaly Detection Techniques

Anomaly detection is a problem to find patterns in the data that behaves abnormally. These patterns are generally referred as outliers, exceptions, surprises in a dataset. Outlier and anomaly are mostly used in the same context in anomaly detection [6]. Anomaly detection is a common problem in data science and has wide variety of usage areas like credit card fraud detection, intrusion detection for cyber security, insurance and healtcare in medical areas etc. In order to understand anomaly detection, definition of anomaly and types of anomaly should be done.

a) What is Anomaly?: Anomalies are patterns in the data that acts abnormally. In Figure 4, visualization of an example of anomaly is shown: Points that are too far away from the normal regions O2 and O3 are considered anomalies. Anomalies may occur in the data for various reasons:

- Data may be corrupted or false data may have entered accidentally in the dataset.
- Anomalies can also occur in the dataset intentionally, credit card fraud, cyber-intrusion etc are the examples of anomalies that can be found in the dataset to act illegally.

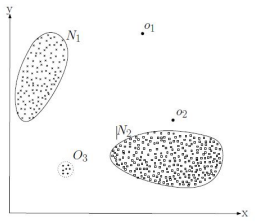

Fig. 4. Point anomaly example

Common characteristic of all anomalies are that they are interesting to the analyst [6].

b) Aspects of Anomaly Detection: In order to understand the anomaly and detect it, nature of the input data, type of anomalies, data labels and the output of the anomaly detection should be discussed first.

Nature of the input data is required to be known in order to determine which anomaly detection technique is going to be used. First type of the data needs to be determined. For example, if the data is categorical data and analyst chooses to use DBSCAN or K-means as clustering technique, result will be wrong.

Type of anomaly should be observed in order to make conclusions about what happened in the data. Types of anomalies are summarized below:

- Point anomalies: If a specific data instance in the dataset is acting abnormally compared to the whole dataset, that instance is considered as point anomaly. An example point anomaly is shown in Figure 3.
- Contextual anomalies: If a specific data instance in the dataset is abnormal in its contextual meaning, that instance considered as an contextual anomaly. For example, in an monthly temperature graph, 10°C is normal for December but abnormal for June so if that is the case data instance in June is considered as contextual anomaly.
- Collective anomalies: In collective anomalies, instead of a single data instance in the dataset, a collection of data instances act abnormally. For example, in an EKG diagram, if the same low value exists for a long time that means there is dysrhythmia occured in that person's hearth.

Data labels are another important aspect of the anomaly detection. Based on the extent to which the labels are available, anomaly detection techniques can operate in one of the following three modes:

- Supervised Anomaly Detection: Techniques that use supervised anomaly detection assumes that all data instances are labeled in the dataset. All data instances are compared to a model to check if it is an anomaly or not. Two issues in supervised anomaly detection is anomalous instances are too few compared to the normal instances and getting the labels of anomalies in the dataset is challenging. It is very similar to predictive models used in machine learning.

- Semi-Supervised Anomaly Detection: Techniques that use assumes that only normal instances are labeled in the dataset and abnormal instances are unlabeled. Difference on semi-supervised and supervised is in supervised detection model is designed for all data instances and compared to whole dataset but in semi-supervised detection model is designed for normal labeled data instances and model is compared to unlabeled data instances in order to find the anomalies in the dataset.
- Unsupervised Anomaly Detection: Techniques that use unsupervised anomaly detection do not require a training data [6]. Unsupervised anomaly detection techniques assume that count of normal instances are superior to the abnormal instances thus it can be applied in almost every dataset. However, if the assumption is wrong and normal instances are not superior to the abnormal instances, false results will occur.

Output of anomaly can be visualized in two ways, scoring the anomalous instances or labeling them as normal or abnormal. If the analyst chose to visualize them using scoring, a threshold can serve as the determining point in the dataset and looking at the instances which are above or below the threshold will show if the instance is anomalous or not. If the analyst chose to label the data instances as normal or abnormal, looking at the labels of data instances, it can be determined if an instance is normal or not.

IV. Methodology

In this assignment, initial dataset consists of 4 columns and 1839 rows. Before trying to detect anomalies in the dataset, prior analysis should help on understanding the data. In order to analyze the given dataset, python programming language is used.

Initial dataset sample is shown below in Figure 5.

	A	B	C
1	Line	BoardingTime	PassengerCount
2	KC06	2019-12-18T00:00	1
3	LF09	2019-12-18T00:00	3
4	TC93	2019-12-18T00:00	1
5	TL94	2019-12-18T00:00	1
6	TL94	2019-12-18T01:00	1
7	VF01	2019-12-18T03:30	4
8	AC03	2019-12-18T05:30	7
9	AF04	2019-12-18T05:30	1
10	DC15	2019-12-18T05:30	33

Fig. 5. Initial Dataset

In initial dataset, first BoardingTime column is split from letter T into two new columns Day and Time. After BoardingTime is split into new columns, it is no longer required so dropped from the dataset and new dataset dfnew is created. Sample of dfnew is shown in Figure 6.

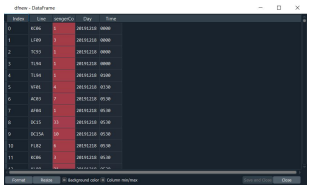

Fig. 6. Initial Dataset

After the dataset is prepared, it is time to choose which type of clustering is going to be used in this assignment. In order to decide which technique is going to be used, data types in the dataset needed to be checked. In Figure 7 data types of the dataset is shown.

```
Line             object
PassengerCount    int64
Day              object
Time             object
dtype: object
```

Fig. 7. Data Types

According to Figure 7, dataset contains mixed data. DBSCAN method only works for numerical data so it is not suitable for this kind of dataset. BIRCH technique requires more numbers of data so it is not suitable too. Hence as explained before, K-modes clustering, which is altered version of K-means clustering for mixed data types is chosen as the proper technique for this assignment.

In order to apply K-modes, it is needed to determine how many clusters would be the best to apply in algorithm. In order to find the optimal number of clusters, Elbow method is used. In Figure 8, output of the elbow method is shown. After observing Figure 8, there is an elbow point if number of

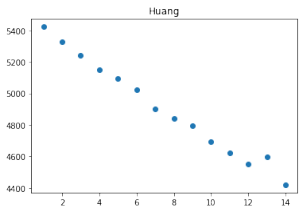

Fig. 8. Elbow Method

clusters is chosen as 13. So number of clusters is determined as 13 and used in the rest of the analyzing.

V. Results and Discussion

Before applying clustering to detect anomalies, dataset is analyzed first. Total number of passengers that used bus in the given day is 318322.

- Total number of passengers that used bus in the given day is 318322.
- There are 51 lines in service according to the dataset.
- KL08 has the most passenger count in the dataset in time 15.30. It is shown in Figure 9.

	Line	PassengerCount	Day	Time	clusterspredicted
1002	KL08	990	20191218	1530	0
997	KC06	908	20191218	1530	0
1257	KL08	882	20191218	1800	0
1155	KL08	877	20191218	1700	0
1206	KL08	863	20191218	1730	3

Fig. 9. Highest Passenger Count

After observing this interesting information, it is time to cluster the data and discuss the results. First lets look at time-passengercount-cluster plot shown in Figure 10. In Figure

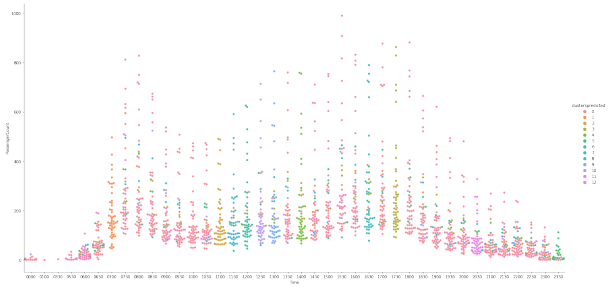

Fig. 10. Time-PassengerCount-Cluster

10, looking at the time interval, at 15.30 highest counts of passengers are observed. Considering working hours are generally between hours 8.00-17.00, it is normal to observe an increase in passenger counts between 6.30-8.00 and decrease after 18.00. However, looking at the figure, having highest counts in 15.30 looks like an anomaly at first. Considering there is no Covid-19 pandemic in 2018 and 15.30 is a suitable for high school students to finish their lessons and use the bus to their way home, resulting as in context there is no anomaly in that time interval.

It is time to look at the line-passengercount-cluster plot to check if there is an anomaly or not. Plot is shown in Figure 11. From the Figure 11, it looks like Line UC32 has an anomaly.

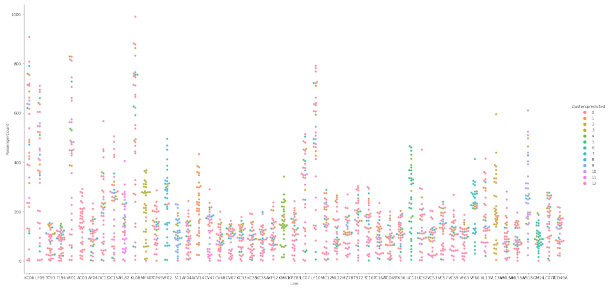

Fig. 11. Line-PassengerCount-Cluster

To further investigate this we need to look at line-time-clusters graph. It is shown in Figure 12. According to Figure 12, our

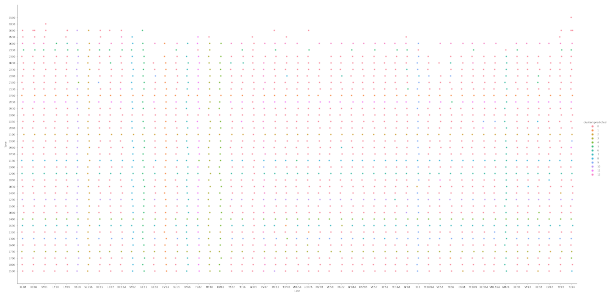

Fig. 12. Line-Time-Cluster

prior detection is correct. At time 19.00 line UC32 makes a peak in passenger count. After searching for route of UC32, conclusion is there is no alternative route for the stops of UC32 on that time interval so passengercount of line UC32 is increasing according to that. Further investigation on relations between Figure 10,11 and 12 are going to be applied as future research.

To check if there is an anomaly on time-cluster or line-cluster, graphs of each are analyzed. In Figures 13 and 14 they are shown.

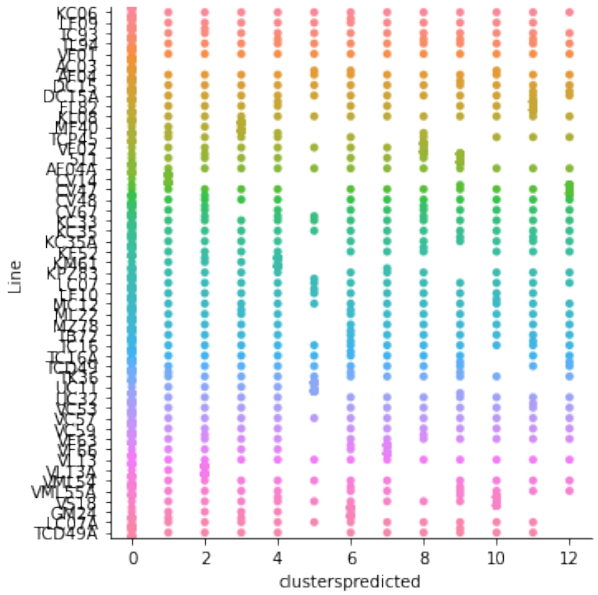

Fig. 13. Line-Cluster

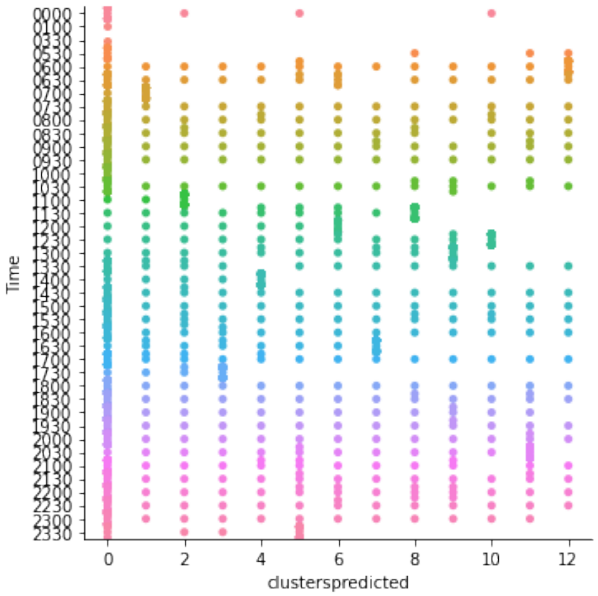

Fig. 14. Time-Cluster

There are no anomalies detecting after analyzing Figure 13 and 14.

VI. Conclusion

In this assignment, passenger data in 18-12-2018 is analyzed. It is observed that the most busy line that day was KL08 and cause of that is the route of KL08 visits some of the critical locations in the city. People use public transport more before and after working hours. According to the news on the internet, there was heavy rain on that week. According to those news, heavy rain may cause people to prefer public transport instead of going to work with their personel vehicles. Since there is no information about the other days, it may be a false assumption but if dataset of near dates are analyzed, there may be a connection.

As a result, this assignment showed that passenger behaviour is determined according to working conditions, available routes and study hours in the city.

References

[1] M. A. Dalal and N. D. Harale, "A survey on clustering in data mining," in *Proceedings of the International Conference & Workshop on Emerging Trends in Technology*, ser. ICWET '11. Mumbai, Maharashtra, India: Association for Computing Machinery, Feb. 2011, pp. 559–562. [Online]. Available: https://doi.org/10.1145/1980022.1980143

[2] Garima, H. Gulati, and P. Singh, "Clustering techniques in data mining: A comparison," in *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, Mar. 2015, pp. 410–415.

[3] M. S. Ejaz, M. A. Hossain, A. Matin, and M. T. Ahmed, "Performance Comparison of Partition Based Clustering Algorithms on Iris Image Preprocessing," in *2017 2nd International Conference on Electrical & Electronic Engineering (ICEEE)*. Rajshahi: IEEE, Dec. 2017, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/8412898/

[4] I. A. Venkatkumar and S. J. K. Shardaben, "Comparative study of data mining clustering algorithms," in *2016 International Conference on Data Science and Engineering (ICDSE)*. Cochin, India: IEEE, Aug. 2016, pp. 1–7. [Online]. Available: http://ieeexplore.ieee.org/document/7823946/

[5] S. S. Khan and S. Kant, "Computation of Initial Modes for K-modes Clustering Algorithm Using Evidence Accumulation," in *IJCAI*, 2007.

[6] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, Jul. 2009. [Online]. Available: https://dl.acm.org/doi/10.1145/1541880.1541882