K-Means Clustering

1st Oguz Kaan Tuna
Hochschule Hamm-Lippstadt
Electronic Engineering
6th Semester
oguz-kaan.tuna@stud.hshl.de

***Abstract*—Clustering is a technique, in which the task is to partition a dataset into groups, called clusters, and to group similar objects into distinct clusters, meaning that the data points, which are similar to each other are in one cluster and the other data points, which are dissimiler to them, to another cluster. This technique is being used in many fields, such as in data mining, pattern recognition or image analysis [1].**

***Index Terms*—**

I. Introduction

Machine learning is an operation, which gives the systems the capability to improve and learn by itself with the help of algorithms, on the data they consume. There are various kinds of machine learning algorithms in the field of data science, which extracts information from a data set to create a model. These algorithms can be categorized in two sections, called Supervised and Unsupervised. In Supervised machine learning algorithms, the algorithm is trained by a dataset, which is labeled priorly. Meaning, it learns to identify an object by memorizing them first. The name supervised in this context is ment to be for instance a data scientist, who guides and teaches the algorithm for which outcomes it should deliver.

In the Unsupervised machine learning algorithms however, the algorithm has to go through the data itself, which means there are no instructor to teach the algorithm. With this approach the algorithm can explore or discover patterns and extract information from the data on its own. Instread of memorizing, the algorithm would learn to identify objects by observing and comparing them so it can seperate the objects into groups and label each specific group. One of the braches of unsupervised learning is clustering. This paper will highlight the idea behind clustering, more specificly the K-means Clustering. The concept and its applications alongside with the advantages and disadvantages of k-means Clustering will be portrayed.

II. Unsupervised Machine Learning Algorithms

As mentioned before, unsupervised learning is an approach, in which the goal of the algorithm is to model or distribute the data, so it can learn more about the data. Biggest difference in Unsupervised learning with supervised learning is, it learns to identify complex processes or patterns without a help and guidance from a human. This leads to a major challenge in this technique. Since unsupervised learning algorithms are used for data without any label information, there is no possible way for the user to know if the output is correct as it should be [2]. As a result of this issue, the evaluation of whether the algorithm learned anything valuable, becomes hard. Mostly the single solution for this problem is to review the output manually. For that reason one of the main application of unsupervised learning algorithms are in experimental manners, such as anomaly detection. Additionally a further extension of this method is called clustering.

III. Clustering

Clustering is a method of seperating data points into a number of groups, called clusters,in which the data points similar to each other are in the same group and those data points, which are dissimilar, are in another. In summary it is a technique of dividing data with similar traits and assigning them to specific clusters. Afterwards these cluster groups will get a number,called cluster ID. These cluster ID's can be then used by machine learning systems for simplifying extensive datasets. This approach is mostly used for statistical data analysis, which is applied in different areas, such as image analysis, recognition of patterns, data mining and of course machine learning. To solve various issues, there are various types of clustering. These can be named as:

- Hierarchical Clustering
- Partitional Clustering
- Distribution-based Clustering
- Density-based Clustering
- Constraint-based
- Fuzzy Clustering

IV. K-Means Clustering

K-Means Clustering is one of the most commonly used and also simplest unsupervised learning algorithms for clustering, Fig.1. With this algorithm, it is possible to arrange a dataset with a fixed prearranged number of clusters. The "k" in k-means clustering stands for that predetermined number of clusters. It is also the main task in this clustering algorithm. Choosing the right value for "k" plays a huge role, if choosed randomly, the result could be satisfying but it is also possible to affect the model performance in an unpleasant way , in case the value is wrongly choosen. Therefore there are several methods developed, in order to select a right value for "k".

But the mainly used approaches are called the Elbow Method and the Silhouette Method.

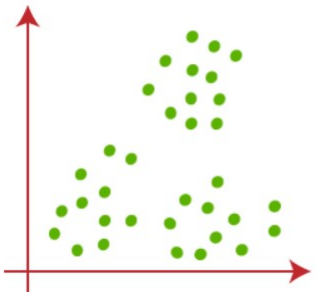

Fig. 1. A data set without the process of clustering.

- Elbow Method

The Elbow Method is one of the most well-known approaches to determine the optimum value for "k". The main process is to run the k-means for various numbers of times with different amount of clusters and decide which is the fitting number of clusters. The idea is, since the number of clusters are increasing the differences between the various clusters are decreasing. But at the same time the differences between the elements inside the clusters are also increasing. The purpose is to find the point, where the elements in a cluster are as homogeneous as possible and the other clusters are as different as possible to each other. Therefore the proper distance between the elements and the center of that cluster, in which the elements are located, should be as small as possible.

- Silhouette Method

The Silhouette method can graphically represent how well an element has been classified. This is done by measuring how similar an item is to ist own cluster in comparison to others. The silhouette value can be between -1 and 1. A higher value signifies that the item is properly matched to its cluster, while beeing fairly not identical to remaining clusters. However if most of the points have largely lower values, this would indicate that the clustering arrangement has less or more clusters than needed.

A. *k-means Algorithm*

$Directk - means$

Initialize k prototypes ($w1$, . . . , w_k) such that $w_j = i_l, j \in \{1, .., k\}, l \in \{1, ..., n\}$
Each cluster C_j is associated with prototype w_j
Repeat
for each input vector $i_{l,}$ where $l \in \{1, ..., n\}$, do
Assign i_l to the cluster C_j* with nearest prototype w_{j*} (i.e.,$|i_l - wj*| \leq |i_l - w_j|, j \in \{1, ..., k\}$)

for each cluster C_j, where $j \in \{1, ..., k\}, do$

Update the prototype w_j to be the centroid of all samples currently in C_j, so that $w_j = \sum_{il \in C_j} i_l / C_j$

Compute the error function:

$$E = \sum_{j=1}^{k} \sum_{i_l \in C_j} |i_l - w_j|^2$$

$Until$ E does not change significantly or cluster membership no longer changes.

The algorithm above shows how the quality of the clustering is determined in the k-means clustering algorithm with a error function. The C_j represents the j_{th} cluster, which has a value of a disjoint subset of input patterns. And the characteristics of the clustering is determined by the error function.

The algorithm is very straightforward and an easy way to cluster a given dataset. These are the steps needed to be performed:

1) Determining the value of "k"

In order to find the groups or clusters, the number of clusters must first be defined. As mentioned prevously, there are several ways to find the optimum number for "k".

2) Creation of random points, "centroids"

In the second step, the initial cluster centroids are then determined. Centroids are random data points, which represents the clusters center. It does not have to be a member of the dataset.

3) Assigning points to the clusters

The distance from the first point to each of the cluster centroids is now measured. The point, that is closest to one cluster centroid is then assigned to that cluster. This is repeated for all the other points. All points are then initially assigned to a group.

4) Calculation of the mean for each cluster

In the fourth step the centroids of the clusters will be recalculated. This will be done by taking the mean of all data points assigned to all cluster. These mean values are the new centroids of the clusters. Therefore tThe "means" in the K-means stands for the averaging of the data and assigning it as the new centroid.

5) Repetition of Steps 3 and 4

The steps of 3 and 4 will be repeated until the centroinds and the cluster division can no longer be changed.If the clusters do not change any more in one iteration, the process will be finished, Fig.2.

B. *Distance Measurements*

Since in k-means algorithm the distance between the points are calculated in order to assign thmen to a centroid with the gathered data, the distance calculation is also playing an important role [3]. Therefore there are several aprroaches developed to calculate the distance between two points, which have different effects on the clustering model. In order to pick the fitting method for a model, some aspects should be noted, for instance the dimension of the data. Some of the frequently used distance measurements are:

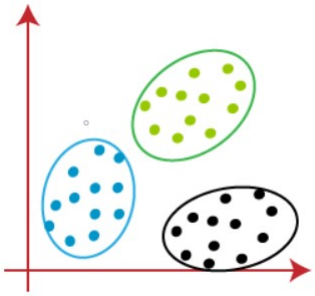

Fig. 2. After k-means is applied.

- City Block
- Euclidean Distance
- Correlation Distance
- Cosine Distance

C. Limitations and Advantages

K-means is overall a very useful science tool in many fields. Since there is only one parameter which needs to be defined, it can be used very effortlessly. Besides beeing easy to implement, it works very effectively with large data sets, which makes it able to deal with massive amounts of data. Also since it is beeing widely used, compared to others algorithms, there are big collections of use cases and implementation in many areas and disciplines. Nonetheless, there are few downsides of this approach, that needs to be pointed as well.

1) Advantages :

- Simplicity

The main phase for the k-means clustering algorithm consists only two steps. These are the assignment of the cluster numbers, ”k”, and creation of centroids. In case a learning algorithm needed, which needs to handle large data sets, k-means can be an appropriate solution.

- Availability and Speed

Since k-means clustering is widely used, most of the machine learning applications can offer the implementation of k-means, such as scikit-learn. Besides beeing widely avaible, k-means is also mostly faster at clusterin compared to other clustering algorithms [4].

2) Limitations :

- Initialization

As mentioned previously, there are no specified initialization opportunities for the centroids. And therefore in case the centroids are selected randomly, there will be variety of clustering models based on the initialized points, Fig. 3, 4.

To solve this complication, the k-means++ algorithm is used. This algorithm is run prior to k-means, in order to determine the most appropriate centroids for the clustering. Some machine learning llibraries, such as scikit-learn, employs k++ by default.

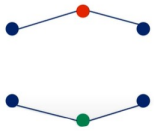

Fig. 3. Despite clustering the points on the right and on the left as one cluster would be appropriate, the k-means algorithm would cluster the points above and the points on the bottom as one cluster, since the centroids are choosen randomly.

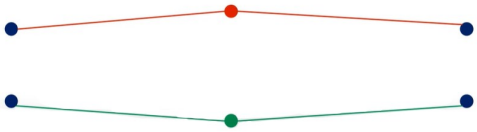

Fig. 4. The same situation as Fig. 3, however this case widely spread points. But as the centroids are the same, the identical cluster model would be encountered.

- Determining the Value for k

As previously referred, defining the value for ”k” plays a vital role in the overall model, reasoned by the fact that the results can be seriously affected by this decision. However there are some approaches avaible, which assists the algorithm to find the suitable values for k, such as the Elbow Method.

- Sensitive to Outliers

Another major problem is that k-means is sensitive to outliers. With outlier is ment, a single point, which is further away from the other points. In this case it will be placed in its own one point cluster. The remedy might be removing those outliers, prior to clustering. Alternatively if one point clusters are spotted, removing them and clustering again can also be an option.

V. Applications

A. Classification of Network Traffic

As the classification of network traffics, based on the port and payload analysis becomes more difficult with time, since the P2P applications are starting to use dynamic port numbers and various encryptions to avoid detection, it becomes harder to understand which types of traffic is coming to a website. In particular, which traffic is spam or which traffic is coming from bots. It is vital to know where the trffic is coming from, for instance in case the traffic is coming with bad intentions, it must be blocked.

In this field clustering algorithms can be used as an alternative, by exploiting the distinct charecteristics of the

application, when they communicate on the network [5]. A research documentation from the University of Calgary [5], shows that by applying k-means clustering algorithm up to 75 % accuracy can be achieved and thanks to its fast model building option, compared to other algorithms k-means is more suitible for this application. It also shown, that clustering techniques work for other applications as well, such as P2P file-sharing or file transfers.

B. Document Classification

Clusternig algorithms in documentations were introduced in order to collect important information in a cluster in an effective and fast manner. The k-means is widely used to cluster documents in multiple categories based on the topics, contents and the text of the documents. With dividing a document in various clusters, people, who are interested in some of the points of a larger documentation, can effortlessly read and gather information from those parts of the documment in which they are interested [6] There are several tools, that carries out the k-means algorithm for this purpose, such as WEKA (Waikato Environment for Knowledge Analysis). It is a free software, which contains visualization tools and different algorithms for data analysis and modeling.

C. Image Segmentation

Image segmentation is a process of dividing an image to several segments. The objective of segmenting an image is make it more easier to analyze. It is mosly used for creating boundaries or locating objects. Image segmentation is beeing used in many fields, such as in healthcare. For instance, since cancer is even with todays technology still a fatal illness, beeing able to detect cancer cells as early as possible would perhaps be a life or death situation. Image segmentation techniques are making importans impacts in this context, as the shape of the cancer cells can determine the type of the cancer. Besides in healthcare, others application areas for image segmentation can be named as traffic control systems, self driving cars or locating objects in sattelite images. K-means is beeing used for classifying the pixels based on the pixels with high similarities or high contrast between the regions.

VI. A Framework Example

As previously mentioned, data mining is one of the application fields of k-means clustering. The goal of data mining is to process data in order to find valuable patterns, in which the pattern can help to decide future trends from larger datasets. The needed information can be collected via several ways, such as decision trees, nearest neighbour method or artificial intelligence. Currently data mining approaches are being widely used in different areas, namely marketing, finance, telecommunication or medical data analysis. With advancing technology, the medical field is becoming a key area for applying data mining techniques. One of the fields, where it can be applied is in diabetes data, which affects millions of people throughout the world.

To use clustering method, a framework based on cloud for diagnosing diabetes was proposed [7]. The expected framework is consisting of three main stages to manage a diabetes dataset, as it can be seen in Fig.5. The Hadoop Distributed File System (HDFS) is a storage medium, which enables the distribution of large data to k-means MapReduce function. The MapReduce function is responsible with processing the extensive dataset. This function includes the Mapper and Reducer functions inside it. The mappers job is to divide the large data into smaller parts. It then arranges different information as a transitional dataset. After the reducer further classifies the data, R-Data Visualisation uses this output and creates a statistical record of diabetes analysis based on the various attributes of a person, such as age, gender and so on. The report is then accesible by the respected scientists or doctors.

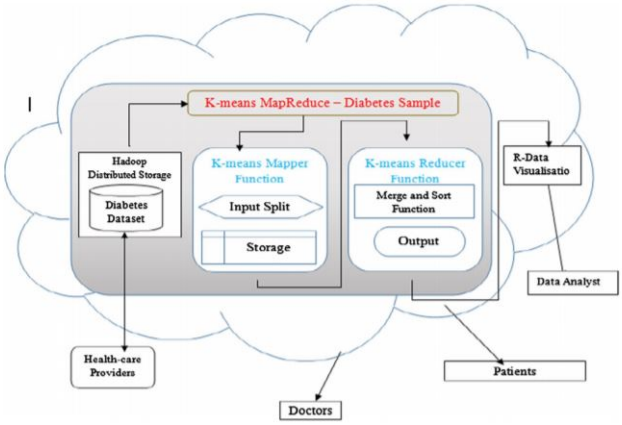

Fig. 5. A cloud architecture based on k-means clustering.

VII. Scikit-learn Library

Scikit-learn is a machine learning library for Python programming language, which features different algorithms in classification, regression and clustering.

A. Simple example of Using scikit-learn for k-means

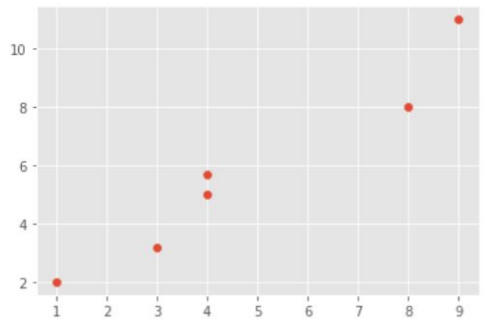

Fig. 6. The unprocessed data set.

In order to use k-means algorithm with the scikit-learn, the needed libraries must be imported first, to display the output for this specific example, Fig.6. We create 6 points and display it accordingly. In Fig. 7, we define how many clusters we want to have and create two variables, ”labels” and ”centroids” to display the results after the clustering process has finished. The final output is displayed in Fig. 8. The ”x” markings are representing the centroids of each cluster, where the different colors the corresponding clusters. The full code for this example can be seen in the apendix.

```
In [3]: X = np.array([[1, 2],
                      [4, 5],
                      [3, 3.2],
                      [8, 8],
                      [4, 5.7],
                      [9, 11]])

In [4]: kmeans = KMeans(n_clusters=2)
        kmeans.fit(X)

        centroids = kmeans.cluster_centers_
        labels = kmeans.labels_
```

Fig. 7. Defining the cluster numbers.

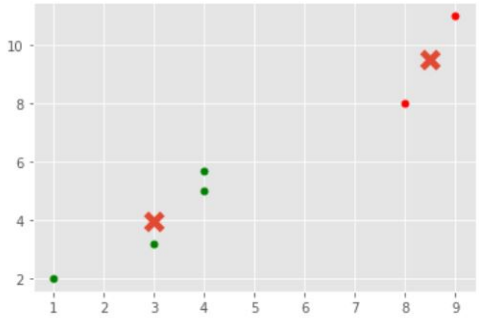

Fig. 8. The final output after clustering process.

VIII. Comparison to other algorithms

A. Comparison 1

B. Comparison 2

Appendix

IX. First References

[1] (Madhulatha, 2012)
[2] (Müller & Guido, 2016)
[3] (Bora, Jyoti, Gupta, & Kumar, 2014)
[4] (McInnes, Healy, & Astels, 2016)
[5] (Erman, Arlitt, & Mahanti, 2006)
[6] (Balabantaray, Sarma, & Jha, 2015)
[7] (Shakeel, Baskar, Dhulipala, & Jaber, 2018)

References

Balabantaray, R. C., Sarma, C., & Jha, M. (2015). Document clustering using k-means and k-medoids. *arXiv preprint arXiv:1502.07938*.

Bora, M., Jyoti, D., Gupta, D., & Kumar, A. (2014). Effect of different distance measures on the performance of k-means algorithm: an experimental study in matlab. *arXiv preprint arXiv:1405.7471*.

Erman, J., Arlitt, M., & Mahanti, A. (2006). Traffic classification using clustering algorithms. In *Proceedings of the 2006 sigcomm workshop on mining network data* (pp. 281–286).

Madhulatha, T. S. (2012). An overview on clustering methods. *arXiv preprint arXiv:1205.1117*.

McInnes, L., Healy, J., & Astels, S. (2016). Benchmarking performance and scaling of python clustering algorithms. *hdbscan. readthedocs. io*.

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with python: a guide for data scientists*. ” O’Reilly Media, Inc.”.

Shakeel, P. M., Baskar, S., Dhulipala, V. S., & Jaber, M. M. (2018). Cloud based framework for diagnosis of diabetes mellitus using k-means clustering. *Health information science and systems*, *6*(1), 1–7.