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

A. Advantages

B. Disadvantages

C. Applications

V. Comparison to other algorithms

A. Comparison 1

B. Comparison 2

Acknowledgment

VI. First References

[1] (Madhulatha, 2012)
[2] (Müller & Guido, 2016)

References

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

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