Study of Unsupervised Machine Learning Using Clustering Algorithms

*Alekhya Roy, Ritayu Shetty, Vikram Sharma
1 Mukesh Patel School of Technology Management and Engineering
NMIMS(Deemed-to-be) University, Maharashtra, India
roy.alekhya@gmail.com

Mukesh Patel School of Technology Management and Engineering
NMIMS(Deemed-to-be) University, Maharashtra, India
ritayushetty12228@gmail.com

Mukesh Patel School of Technology Management and Engineering
NMIMS(Deemed-to-be) University, Maharashtra, India
vikram.sharma.main@gmail.com

Abstract

With the rising popularity of AI, taking a look at its deficiencies is significant in perceiving how well AI is applicable. Is it conceivable to apply the grouping with a small dataset? This paper comprises of a writing study, a review and an examination. It examines how two distinct AI calculations DBSCAN(Density-Based Spatial Clustering of Applications with Noise) and K-implies run on a dataset. Making a study where we can see genuinely what the vast majority picked and apply grouping with the information from the review to affirm in the event that the bunching has similar examples as what individuals have picked measurably. It was feasible to recognize designs with grouping calculations utilizing a little dataset. The writing concentrates on show models that the two calculations have been utilized effectively. It's feasible to see designs utilizing DBSCAN and K-implies on a little dataset. The size of the dataset isn't really the main perspective to think about, component and boundary choice are both significant as the need might arise to be tuned and tweaked to the information.

1 Introduction

A lot of information and data are being gathered from various gadgets like telephones, PCs, vehicles, GPS and a wide range of associated gadgets. In the present period with a lot of information, the time taken to compute is increased, and this is where machine learning and AI comes right into it, to assist with handling enormous information in a sensible measure of time. AI is a subfield of man-made consciousness, which is an area of software engineering that underscores the production of keen machines or projects that work and respond like people. Objectives of AI can be to program PCs to utilize model information or previous encounters to take care of a given issue. AI can assist us with understanding the design of information and squeezed that information into models that can be perceived and used by people.
The idea of Machine learning was first characterized in 1959 by Arthur Samuel as the field of study that enables PCs to learn without being unequivocally modified [11]. With the ongoing handling power and headways in the field it makes AI more feasible and pertinent in present day days [1]. With prospects to turn up devoted servers from organizations, for example, Amazon and lease superior execution equipment makes it simple for even people to work with cutting edge and computational weighty projects and enormous datasets. AI can be classified into three unique branches which are managed learning, unaided learning and support learning.
Administered AI is the point at which the part is seeing from information and result information. The info information necessities to incorporate names (characterized as right information) [10]. The objective is to comprehend the planning of how they connect with each other. This incorporates themes like relapse, expectation, and order.
Unaided AI shares much for all intents and purpose with exploratory information investigation and information mining, it's just a perception of the information. There are just information and not accessible result perception. The information is unlabeled so there is no correct in the information. It's confined to understanding what can be gained from the information.
Support AI incorporates a specialist who makes a move contingent upon the circumstance and gets compensated for doing its activities. This learning strategy doesn't have to indicate how the activity ought to be taken care of; the specialist just gets compensated by playing out the right activities. The objective is to have a specialist who does activities accurately from doing experimentation learning in a unique climate [8].

This proposition will examine and be limited to the investigation of the unaided learning branch. It will research how well K-Means and DBSCAN work on little datasets. These are bunching calculations that utilization various methodologies. DBSCAN is thickness based and K-implies is a centroid based calculation, this will make it more intriguing to contrast them and one another. We have characterized a little dataset as a dataset with under 500 perceptions. The dataset utilized is gathered through an overview and zeroing in on people groups preparing propensities. It will research on the off chance that it is conceivable with a dataset with few examples to obtain intriguing outcomes, or on the other hand on the off chance that we want more examples of information to make any important bunches. To check whether it is feasible to make any bunches containing clear groupings of people groups, for example bunches containing individuals with a similar age, orientation and so on.

Utilizing AI for the improvement of projects make them more solid and it goes quicker then, at that point, assuming a human were to foster the program without any preparation [1]. The negative part of utilizing AI is the requirement for information to prepare the program, it's likewise more PC weighty than a customary program. AI can do a ton to work on our reality, it can work on different various fields beyond programming improvement too. A model being financial matters, as [1] makes reference to. Replacement, cost flexibility, pay versatility and more could be improved by utilizing AI. Most seemingly anyway the field that AI can change the most is computerization. Having the part accomplish the work is less expensive and it doesn't require breaks. As [22] notices the AI calculations grasps the idea and utilize the suitable way to some random region. This, thusly, could influence some workspaces in the way that the work obligations can be more mechanized or even totally replaced by self-learning components.

2 Research Questions

In this section, we present the research questions, and explain why we came up with them. We also delineate our goals and objectives, and then what we think the expected outcome will be.

2.1 Research Questions

RQ1: In which domains can DBSCAN and K-means be utilized in?

RQ2: What observations can be made by noticing at the patterns from the unsupervised clustering algorithm DBSCAN on people's training habits using a small dataset?

RQ3: What observations can be made by observing at the patterns from the unsupervised clusteringalgorithm K-means on people's training habits using a small dataset?

2.2 Background

This paper will explore where unsupervised clustering algorithms can be applied through a literature study, this will acquire the information on where such calculations can be applied. The examination is limited to focus on two grouping calculations. These calculations utilize different clustering approaches, DBSCAN is density based and K-implies is a centroid based algorithm, which makes the examination and the grouping more fascinating since getting various clusters will be conceivable. The inspiration for picking DBSCAN is that it can consequently decide the quantity of groups, it can deal with information with commotion/exceptions and it can likewise recognize anomalies while distinguishing bunches [12]. K-implies was picked in view of its wide ubiquity and effortlessness. These calculations will bunch the dataset accumulated from the overview. The constraint of the dataset is set to 500 examples from our study, as our meaning of a little dataset is under 500. This is done in light of the fact that we need to zero in on a dataset with less examples. Research question 2 and 3 will give and present an examination through an investigation of these bunching calculations on the dataset. The worth of this is propelled by the size of the dataset. The objectives with research question 2 and 3 are to check whether it is feasible to sum up the bunches in light of a little dataset. The primary goal is to explore in the event that we can distinguish designs in the bunches. Designs in the bunches can, for instance, be assuming that it is feasible to distinguish what kinds of gatherings exist or to recognize obscure gatherings with comparable propensities. While contrasting DBSCAN and K-implies we need to think about the groups and investigate assuming that they produce comparable outcomes.

2.3 Expectations

From the survey we anticipated at least 100 responses. We expected it would take a limit of about fourteen days to get that number of replies. For the writing question (RQ1) we read papers/articles about what fields K-means and DBSCAN can be utilized in. We hope to track down a variety of fields for the two calculations since both are well known. We additionally hope to obtain definitely various outcomes relying upon what fields the algorithms are utilized in on account of their distinction in tracking down groups. The explanation being that K-implies is centroid based while DBSCAN is density based.
For research question (RQ2) we expected to notice bunches of inconsistent shapes and hence have the option to recognize what gatherings have the most comparative propensities. We additionally expected to distinguish anomalies in the dataset, for example tests of information that have a place with no group. This is fascinating in light of the fact that we can then distinguish examples and see what compels them vary from the rest. For the exploration question 3 (RQ3) we bunch the information with the calculation K-Means. We expect contingent upon the predefined number of bunches to come by altogether different outcomes. In this question we hope to notice the manner in which K-implies chips away at a dataset which will put individuals with comparable highlights in similar groups however depending of the quantity of bunches this can be split to additional separated bunches. The assumption from utilizing the two calculations is to track down various speculations. We accept that the two calculations will have no issue to bunch the given information, what we are uncertain of is on the off chance that we can get anything from checking the groups out. The assumptions we have is that K-means will make exceptionally broad speculations, this thus could make the speculation too vague making the outcome dull. To come by around the outcome being to vague we will characterize a larger number of bunches to obtain an outcome that we can gain something from. DBSCAN will be less broad and give us inconsistencies, we anticipate that outcomes with DBSCAN should be more unambiguous however we are uncertain assuming the information is excessively little to get explicit. This could bring about DBSCAN bunching nearly everything or barely anything. To get around this we will change minpoints and epsilons to generalize. ustering algorithms

In this section, we present the two clustering algorithms that were utilized in the proposal. We give a general show of them and give four perceptions to every calculation. The objective here is to give you a speedy, significant level introductions of the algorithms to all the more effectively track.

2.4 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu in 1996 [6]. The algorithm catches the knowledge that assuming a specific point has a place with a cluster, it ought to likewise be near different places in that cluster. One of the upsides of utilizing DBSCAN is that it can track down inconsistent states of clusters. DBSCAN has been applied in various fields, for example, network traffic classification [5], PC security, for example, malware classification[9], environment studies[5] and anomalydetection[12]. DBSCAN relies upon two boundaries, a positive number epsilon and the base number of focuses called minPoints. At first, all data of interest in the dataset are unassigned. DBSCAN starts by picking an inconsistent piece of information from the dataset that has not been visited. On the off chance that there are more than minPoints focuses, including itself, within the distance of epsilon from point p, then those focuses structure a cluster. Point p is supposed to be a center point and the neighbor focuses within the distance of epsilon are supposed to be straightforwardly reachable from point p. DBSCAN really looks at every one of the new places in the cluster to check whether they also have more than minPoints focuses within a distance of epsilon, on the off chance that that is the situation, DBSCAN grows the cluster by adding them to the cluster, it then, at that point, rehashes this step for the recently added places. At the point when there are no more focuses to add to the cluster, it picks a new inconsistent, unvisited point from the dataset and rehashes the cycle. In the event that the point has not exactly minPoints focuses within the distance of epsilon and has a place with no other cluster, then it's viewed as a noise point.

2.5 K-means

The K-implies algorithm was first referenced by James MacQueen in 1967[14], but the thought began in 1957 by Hugo Steinhaus [23]. The K-implies is a centroid-based clustering algorithm. K-implies algorithm has been applied in different fields, for example, picture division [3], illness expectation [15], network traffic classification [5]. The algorithm parcels n tests of a dataset into a decent number of k disjoint subsets/clusters where each example has a place with one of the k clusters. The worth of k should be predefined. The focuses of the clusters are called centroids and are at first picked arbitrarily from within the subspace. K-implies algorithm works in 2 stages, in the initial step all information focuses are doled out to the cluster with the closest centroid. In the subsequent step, all clusters recalculate and refreshes the centroids area based on the mean of all information focuses appointed to their clusters. These 2 exchanging steps go on until the centroids quit moving.

3 Literature Review

This is a synopsis of the examination papers that were found during the writing study for research question

Anomaly Detection in Temperature Data Using DBSCAN Algorithm [16]

The paper discusses tracking down abnormalities in monthly temperature information. The strategy that was utilized beforehand in finding peculiarities was a factual technique. While contrasting DBSCAN with the measurable strategy, DBSCAN was significantly more precise in tracking down the right peculiarities. The main disadvantage is that the information needs preprocessing to get a decent anomaly location. The issue with the information is that the temperature that it considers an anomaly change from what season it is. To eliminate the occasional element, a z-score is utilized, to sum up, the information by what season it is. Later the information gets standardized by the mean and standard anomaly of the month. The temperature information anomaly recognition examination is utilized in different various fields, for example, administration, general wellbeing, and environmental studies.

Traffic classification using clustering algorithms [5]

This paper discusses how to precisely distinguish and arrange network traffic as per application type. The authors portray that it is a significant component of many network executives' errands, for example, stream need, traffic molding/policing, and symptomatic checking. The authors depict 3 unique methodologies of how to distinguish and order network traffic and discuss the benefits and burdens of a particular methodology. In the primary methodology, the classification of network traffic is based on the mappings of known ports to applications. In the subsequent methodology, the packet payload is broken down to decide if they contain qualities of known applications. The paper zeroed in on a third methodology, where they investigate how to order information based on just transport layer measurements with clustering. They show that cluster investigation can bunch and sort network traffic utilizing just transport layer traffic with DBSCAN and K-Means, utilizing exact Internet trails. The trial results show that both K-Means and DBSCAN function admirably and considerably more rapidly than AutoClass. That's what the outcomes show despite the fact that DBSCAN has lower exactness contrasted with K-Means and AutoClass, DBSCAN delivers better clusters.

A New Approach of Image Segmentation Method Using K-Means and Kernel BasedSubtractive Clustering Methods [3]

The paper discusses how image segmentation is the most important phase in image handling and a proposed machine learning algorithm to make an exact segmentation. The authors utilized clustering as a result of the great benefits, its effortlessness, and its effectiveness. The proposed algorithm is a blend of K-means and kernel-based subtractive strategies. The kernel capability is there to build the effectiveness by changing the pixels from the image into different aspects to additional effectively independent them. K-means is subsequently used to recognize the various sorts of portions in the image.

Brain Tumor Segmentation Using Fuzzy C-Means and K-Means Clustering and Its Area Calculation and Disease Prediction Using Naive-Bayes Algorithm [15]

The paper discusses how with the assistance of machine learning finding brain cancers and the hazard of disease is conceivable. Today the discoveries are finished utilizing magnetic or radiation kinds of outputs which require some investment and are costly. Assuming image segmentation is finished by utilizing K-means and Fuzzy C-means it would both save time and make the interaction less expensive. This is finished by creating Brain MRI images, preprocessing the image, utilizing the two algorithms, and finally anticipating the illness by noticing the segmentation result. For finding mass cancer utilizing K-means is sufficient, if any way there is noise in the MR image, K-means needs preprocessing by sifting the noise. K-means isn't adequate all alone as it doesn't identify exhaustively and for this reason, Fuzzy C-means is subsequently utilized after K-means to get more exact growth shape extraction.

Malware classification based on call graph clustering [9]

In this paper, the specialists investigate the possibilities of call graph-based malware ID and classification. The clustering algorithms utilized in the examinations incorporate different renditions of the k-medoids clustering algorithm, as well as the DBSCAN algorithm. The authors portray that it is attractive to recognize gatherings of malware with solid primary likenesses and depicts that enemy of infection motors can utilize conventional marks, focusing on the common similitudes among tests in a malware family. The distinguishing proof of gatherings of malware with solid underlying likenesses is done by means of clustering algorithms. The end expresses that k-means clustering isn't powerful to find malware families, primarily in light of the

fact that deciding the ideal number of clusters was impractical. The outcomes performed with DBSCAN were fruitful since it was feasible to recognize malware families.

Traffic Anomaly Detection Using K-Means Clustering [17]

This paper centers around network information mining and stream-based anomaly recognition by utilizing the K-means clustering algorithm. K-means is utilized to isolate time spans by contrasting ordinary and irregular clusters. The clustering is finished with 3 highlights, the number of packets, and the number of bytes that permits anomaly location. The third component helps security by checking distinguishing network port sweeps and conveyed attacks. This thus expands the quantity of source-objective parts. Still involving k-means in these elements for the dataset isn't sufficient to track down all peculiarities. By utilizing two preprocessing instruments it's feasible to find all anomalies as long as the quantity of exceptions is little. The main device utilized is doing a distance estimation on hubs that aren't important for a centroid cluster and putting them in the cluster they have a place with. The estimation is finished by estimating what centroid is nearest to the hub. The other strategy is outliers' detection, which does a computation to check whether a hub is excessively far from a typical cluster by a limit it's characterized as an anomaly. Both of these techniques are joined to make the classification an atypical and ordinary way of behaving.

Towards a Hybrid Approach of K-Means and Density-Based Spatial Clustering of Applications with Noise for Image Segmentation [7]

This paper proposes a hybrid technique to deal with image segmentation utilizing what the researchers of the paper call Kmeans-DBSCAN. Due to the high computational intricacy of DBSCAN and the enormous size of image datasets, K-means is applied to decrease the size of image datasets in the proposed approach. The clusters centroids created by K-means are additionally clustered by DBSCAN. The image segmentation results are at last given by combining the consequences of K-means and DBSCAN.

Anomaly detection in onboard-recorded flight data using cluster analysis [12]

This paper presents a technique that assesses flight information and recognizes irregular trips without determining the standards which characterize an anomaly. This is accomplished with cluster analysis strategies. DBSCAN is utilized as the algorithm to cluster the dataset. The dataset utilized in the examination comprises a Digital Flight information Recorder (DFDR) dataset from a global carrier. The creators of the paper presume that the underlying assessment shows that cluster investigation is a promising methodology for distinguishing proof of peculiar departures from installed recorded flight information.

3.1 Literature comparison

In this section we compare the different literature with each other and describe when references are supporting each other, and when they are in conflict with each other.

The analysts in the papers that we have contemplated have all accomplished victories from the DBSCAN algorithm. In the paper[7], a combined variant of K-means and DBSCAN is utilized with the inspiration of the great computational intricacy of DBSCAN and the huge size of image datasets. K-means is applied to decrease the size of image datasets. Different papers that consider DBSCAN, all have little enough dataset to keep away from the adverse consequences of the high computational intricacy. The inspiration for involving DBSCAN in the various papers is principally a direct result of its capacity to track down randomly molded clusters as well as/in mixes with its capacity to recognize exceptions. And furthermore, the way that there is a compelling reason need to determine the number of clusters ahead of time.

As k-means can't distinguish inconsistencies, part [15] generally dislikes it as it upsets the nature of the clusters. The arrangement [15] utilizes is erasing the noise so k-means doesn't decipher the information wrong. Anyway [17] attempts to find noise and irregularities which is something k-means isn't known for. It truly relies upon the information as [17] anomalies are extremely near one another. Consequently [17] characterizes typical and strange clusters with some assistance.

Finding and understanding irregularities could be pivotal in finding security openings in both [5] and [17] attempts to solidify the security by tracking down peculiarities from network information. The two purposes of k-means however have different associate instruments, [5] has DBSCAN for correlation, [17] utilizes two preprocessing strategies distance estimation and exceptions identification

DBSCAN is usually utilized as an anomaly recognition algorithm as both [16] and [12] utilize it. Rather than utilizing an anomaly location algorithm both [16] and [12] utilize DBSCAN a clustering algorithm to track down anomalies.

Utilizing two algorithms is usually finished to get a more precise or better comprehension of the information. [7],[5] and [9] utilize a centroid-based algorithm and DBSCAN. [9] and [5] both utilized K-means separate with DBSCAN for correlations while [7]

involved it for segmentation of the information and later run DBSCAN on each fragment. [9] and [5] obtained an improved outcome

4 Analysis

4.1 Literature study

To address question 1, we gathered 8 papers as our main literature. This was pertinent because it was oriented towards the DBSCAN and K-means implementations and there weren't any problems in finding these papers. Papers are found based on the outputs of Microsoft Academics and Google scholars.

4.2 Experiment

The examination came by exceptionally blended results, it's what we expected yet we figured finding patterns would be more straightforward. While testing on less elements the grouping had more clear examples. To this end I don't sort out comes up in pretty much every grouping run with various determination highlight as most choices are something very similar in the event that you don't work out. It was simpler to distinguish designs when we utilized less all out information, the justification for this can either be a result of the manner in which we handle all out information, where every conceivable choice for each question turns into an aspect/new element. It can likewise be a result of the quantity of potential choices on each inquiry in blend with the couple of quantities of information focuses which bring about numerous varieties in the dataset. In view of this, it's most certainly conceivable to see designs utilizing a smaller dataset.

K-implies and DBSCAN both figured out how to show designs in their bunches, both got comparative re-sults for certain couple of special cases. The principal distinction is that k-implies worked better with many elements while DBSCAN would be wise to grouping results on few highlights. As we ex-pected we got a superior speculation by seeing K-implies and DBSCAN recognized exceptions and more unambiguous bunches. This is comparable outcomes to [5], that reasons that *"The experimental results show that both K-Means and DBSCAN work very well and much more quickly than AutoClass. The results indicate that although DBSCAN has lower accuracy compared to K-Means and AutoClass, DBSCAN produces better clusters."* In the paper [9] it's expressed that K-implies was not accessible to find malware families while DBSCAN was effectively ready to recognize malware families.

From perusing the writing, we anticipated that DBSCAN should over perform against K-implies as it did. The boundary determination is more diligently for DBSCAN anyway as tracking down the right epsilon and minpoints is more enthusiastically than tracking down a lot of bunches of K-implies. This makes K-implies more straightforward to work with while DBSCAN has a superior grouping quality on the off chance that you enjoyed the perfect proportion of exertion with the boundary choice. While utilizing clear cut information there is a scarce difference for epsilon where in the event that you cross a worth marginally there is an enormous distinction. This makes DBSCAN significantly touchier however it's much more adjustable for explicit situations. In the [7] they utilize K-means to section the dataset into more modest fragments, in light of the great computational intricacy of DBSCAN on a to enormous dataset. This demonstrate the way that it might really be valuable at times to diminish the size of a dataset

While it's more straightforward to see designs from utilizing less highlights, finding the right element choice is pivotal to track down the most ideal examples. On the last two tests it's unmistakable how much a particular example changed the outcome, for example Accomplish you work out with an accomplice? as it's essentially a yes and no component it gives high character and loses similitudes. To come by great outcomes bunching for any dataset you must dissect the information and grasp the elements. Having a larger number of information is in every case better compared to having less and you want to have a limit. Meaning the information can be exceptionally negligible regardless have the option to show designs with a decent element determination.

.

5 Conclusion

This theory utilizes two unsupervised clustering algorithms. DBSCAN(Density-Based Spatial Clustering of Applications with Noise) and K-implies on a dataset accumulated through an overview, effectively responding to every one of the inquiries we tried to address The proposition has been directed through a writing study and an examination. The writing concentrate on directed during this postulation planned to address the main examination question, for example in which fields can DBSCAN and K-implies be utilized ready? Also, to get further information about the calculations and in what fields DBSCAN and K-implies where utilized and how they were applied. We found that clustering with DBSCAN and K-implies are applied in fields can and have been applied in various regions. For instance, in malware classification and identification, anomaly detection, image segmentation, public health, traffic classification of network data and climate studies. The assessment on finding plans inside the gatherings made by K-suggests and DBSCAN occurred with mixed results. K-suggests made greater hypothesis which achieved making it possible to see plans yet with two or three anomalies in the gatherings considering the shortfall of idiosyncrasy acknowledgment and that you need to pick the number of bundles ahead of time. While using less components on our dataset the gatherings had a greater extent between each other with extra unsurprising models. While DBSCAN, of course, gained a few harder experiences to get a general gathering end, diminishing the number of components made DBSCAN more unambiguous than K-infers with the capacity to find eccentricities and conflicting shapes. Having a greater size on the dataset will give a more careful depiction from the requests which subsequently gives a more exact clustering to this current reality.

References

[1] Erik Brynjolfsson and Tom Mitchell. What can machine learning do? workforce implications. *Science*, 358(6370):1530–1534, 2017.

[2] JOHN R BUCKLEY. Automation. *Journal of Academic Librarianship*, 20(1):40, 1994.

[3] Nameirakpam Dhanachandra and Yambem Jina Chanu. A new approach of image segmen- tation method using k-means and kernel based subtractive clustering methods. *International Journal of Applied Engineering Research*, 12(20):10458–10464, 2017.

[4] Jennifer G Dy and Carla E Brodley. Feature selection for unsupervised learning. *Journal of machine learning research*, 5(Aug):845–889, 2004.

[5] Jeffrey Erman, Martin F. Arlitt, and Anirban Mahanti. Traffic classification using clusteringalgorithms. In *Proceedings of the 2006 SIGCOMM workshop on Mining network data*, pages 281–286, 2006.

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

[7] Chun Guan, Kevin Kam Fung Yuen, and Qi Chen. Towards a hybrid approach of k-means anddensity-based spatial clustering of applications with noise for image segmentation. In *Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2017IEEE International Conference on*, pages 396–399. IEEE, 2017.

[8] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: Asurvey. *Journal of artificial intelligence research*, 4:237–285, 1996.

[9] Joris Kinable and Orestis Kostakis. Malware classification based on call graph clustering. *Journal in Computer Virology*, 7:233–245, 2011.

[10] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*,160:3–24, 2007.

[11] David Neil Lawrence Levy. *Computer Games I*. Springer, 2011.

[12] L. Li, M. Gariel, R. J. Hansman, and R. Palacios. Anomaly detection in onboard-recorded flight data using cluster analysis. In *2011 IEEE/AIAA 30th Digital Avionics Systems Con- ference*, pages 4A4–1–4A4–11, Oct 2011.

[13] Mark S Litwin and Arlene Fink. *How to measure survey reliability and validity*, volume 7. Sage, 1995.

[14] J. MacQueen. Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 281-297 (1967)., 1967.

[15] Divyani Sanjay Mane and Balasaheb B Gite. Brain tumor segmentation using fuzzy c-meansand k-means clustering and its

area calculation and disease prediction using naive-bayes algo- rithm. *Brain*, 6(11), 2017.

[16] Jenny Matthews and John Trostel. An improved storm cell identification and track- ing (scit) algorithm based on dbscan clustering and jpda tracking methods. *Amer- ican Meteorological Society, Atlanta, GA.[online] Available from: http://ams. confex. com/ams/90annual/techprogram/paper_164442. htm*, 2010.

[17] Gerhard Münz, Sa Li, and Georg Carle. Traffic anomaly detection using k-means clustering. In *GI/ITG Workshop MMBnet*, 2007.

[18] Alexander Novikov, Mikhail Trofimov, and Ivan Oseledets. Exponential machines. *arXiv preprint arXiv:1605.03795*, 2016.

[19] Pandas-documentation.pandas.dataframe . https://pandas.pydata.org/pandas-docs/ stable/generated/pandas.DataFrame.html, March 2018.

[20] sklearn documentation. 4.3.5. encoding categorical features . http://scikit-learn.org/ stable/modules/preprocessing.html#encoding-categorical-features, March 2018.

[21] sklearn documentation. sklearn.preprocessing.onehotencoder . http://scikit-learn.org/ stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html, March 2018.

[22] Nainsi Soni and Manish Dubey. A review of home automation system with speech recognitionand machine learning. *International Journal*, 5(4), 2017.

[23] Hugo Steinhaus. Sur la division des corps matériels en parties. *Bull. Acad. Pol. Sci., Cl. III*,4:801–804, 1957.