

Unsupervised machine learning, an investigation of clustering algorithms on a small dataset.

Alvarez Gonzalez, Pierre
alvarez9549@gmail.com

Forsberg, Fredrik
forsberg.fredrik@hotmail.com



TODO

- Research about how to map categorical values to integer representations - Distance between Categorical Attributes simple matching ✓
- Start experiment ✓
- Comparison between RQ1 and RQ2 ✓
- Summary for each literature text
- Write some more to introduce the literature review
- Literature comparison
- Literature analysis
- Test different parameters results for the experiment
- Document all the results from the experiment
- Thesis analysis
- Thesis conclusion
- Thesis future work
- Thesis abstract
- Re-factor the text and make it deliverable
- Remove TODO list

Abstract

TODO

Contents

1	Introduction	5
2	Clustering algorithms	6
2.1	K-means	6
2.2	Visualizations - K-means	6
2.3	DBSCAN	7
2.4	Visualizations - DBSCAN	7
3	Research Questions	8
3.1	Research Question	8
3.2	Background	8
3.3	Expectations	8
4	Research Method	10
4.1	Literature study	10
4.2	Search Engines	10
4.3	Keywords	10
4.4	Limitations and validity threats	10
4.5	Empirical study	11
4.6	Survey	11
4.7	Design of survey	11
4.8	Experiment	12
4.8.1	Testing environment	12
4.8.2	Preprocessing	12
5	Literature Review	13
6	Result	14
6.1	Literature study	14
6.2	RQ3: In what fields can DBSCAN and K-means be applied?	14
6.3	Survey	16
6.4	Experiment	17
7	Analysis	18
8	Conclusion	19
9	Future Work	20
10	Annexes	21
10.1	Survey Questions	21
10.2	Code	23
	Bibliography	25

1 Introduction

Never before in the history has there been so much information being collected and stored. Large amounts of data and information are being collected from our phones, computers, cars, GPS and all sorts of connected devices. In today's era with large amounts of data, the time taken to compute is increased, and this is where Machine Learning comes into action, to help process large data in a reasonable amount of time. Machine learning is a subfield of artificial intelligence, which is an area of computer science that emphasizes to create intelligent machines, intelligent machines can be machines or programs that work and react like humans. Goals of machine learning can be to program computers to use example data or past experience to solve a given problem. Machine learning can help us to understand the structure of data and fit that data into models that can be understood and utilized by humans.

Machine learning was first defined in 1959 by Arthur Samuel as the field of study that gives computers the ability to learn without being explicitly programmed.[10] With the current processing power and advancements in the field[2] does make machine learning more realizable and applicable in modern days. With possibilities to spin up dedicated servers from companies such as Amazon and rent high-performance hardware makes it easy for even individuals to work with advanced and computational heavy programs and large datasets. Machine learning can be categorized in three different branches. Supervised learning, Unsupervised learning and Reinforcement learning.

Supervised machine learning is when the component is observing from input and output data. The input data needs to include labels(defined as correct data)[9]. The goal is to understand the mapping of how they relate to one another. This includes topics such as regression, prediction, and classification.


Reinforcement machine learning includes an agent who takes action depending on the situation and gets rewarded for doing its actions. This learning method doesn't need to specify how the action should be handled, the agent only gets rewarded by performing the correct output. The goal is to have an agent who does actions correctly from doing trial and error learning in a dynamic environment interactivity.[8] This learning approach is different from classical statical theory.

Unsupervised machine learning has much in common with exploratory data analysis and data mining. Only an observation from the data. There are no input and output observations rather an observation on variables and vectors.[1] The data is unlabeled(no correct answer) so there is no right or wrong in the data. It's restricted to the goal in understating what can be learned from the data.



This thesis will investigate and restrict the study to the unsupervised learning branch. The thesis will investigate how well K-Means and DBSCAN work on small datasets. It will investigate if it is possible with a dataset with few samples to get some interesting results, or if we need more samples of data to create any valuable clusters. To see if it is possible to create any clusters containing clear groupings of peoples, e.g. groups containing people with the same age, gender etc. The dataset used is collected through a survey and focusing on peoples training habits.

Using machine learning for the development of programs make them more reliable and it goes faster then if a human were to develop the program from scratch.[2] The negative aspect by using machine learning is, the need of data to train the program and it's more computer heavy than a regular program. But what can machine learning do to improve our world? In fact it can improve various different fields besides software development. An example is *economics* like [2] mention Substitution, Price elasticity, Income elasticity and more could be improved by using machine learning. But arguably the field machine learning can change the most is automation. Having the component do the work is cheaper and it doesn't need breaks. Like [18] mentions, that machine learning algorithms understands the concept and to use the appropriate manner to any given area. This, in turn, could affect some work areas in the way that the work duties can be more automated or even completely replaced by self-learning components.[3]


2 Clustering algorithms

In this chapter, we introduce the two algorithms that we used in our thesis. We give a high level presentations of them and provide 4 visualizations for each algorithm. The goal here is to give the reader a quick, high level presentations of the algorithms to easier follow along. 

2.1 K-means

K-means algorithm was first mentioned by James MacQueen in 1967[12], however the idea originates back 1957 by Hugo Steinhaus[19]. K-means is a centroid-based unsupervised learning algorithm. The algorithm partitioning n samples of a data set into a fixed number of k disjoint subsets/clusters where each sample belongs to one of the k clusters. The value of K must be predefined. The centers of the clusters are called centroids and are initially chosen randomly from within the subspace. K-means works in 2 steps, in the first step all data point are assigned to the cluster with the nearest centroid. In the second step, all clusters recalculate and updates the centroids location based on the mean of all data point assigned to their ers. These 2 alternating steps continue until the centroids stop moving. 

2.2 Visualizations - K-means

This visualizations show how K-means clusters on 4 different dummy datasets, generated by sklearn. The data set contains for 1000 data points and 2 feature each. There are 2 clusters predefined. 

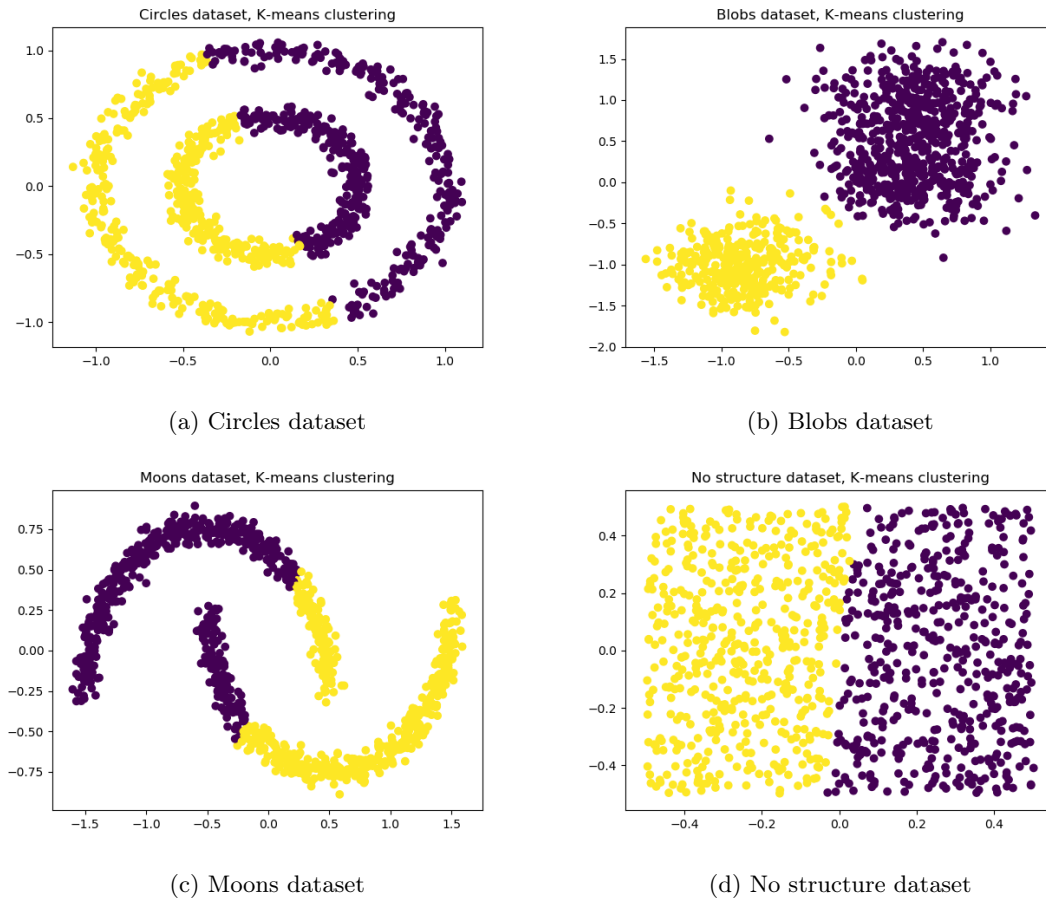


Figure 1: K-means visualization

2.3 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.[7] The algorithm captures the insight that if a particular point belongs to a cluster, it should also be close to other points in that cluster. One of the advantages of a centroid-based clustering algorithm is that it can find arbitrary shapes of clusters. DBSCAN depends on two parameters, a positive number epsilon and the minimum number of points called minPoints. Initially, all data points in the dataset are unassigned. DBSCAN begin by pick an arbitrary data point from the dataset that has not been visited. If there are more then minPoints points, including itself, within the distance of epsilon from point p, then those points form a cluster. Point p is said to be a core point and the neighbor points within the distance of epsilon are said to be directly reachable from point p. DBSCAN check all of the new points in the cluster to see if they too have more than minPoints points within a distance of epsilon, if that is the case, DBSCAN expands the cluster by adding them to the cluster and repeat this step for the newly added points. When there are no more points to add to the cluster, we pick a new arbitrary, unvisited point from the dataset and repeat the process. If the point has less than minPoints points within the distance of epsilon and does not belong to any other cluster, then it's considered a noise point.

2.4 Visualizations - DBSCAN

This visualizations show how DBSCAN clusters on 4 different dummy datasets, generated by sklearn. The data set contains for 1000 data points and 2 features. $Epsilon=0.5$ and $minPoints=5$

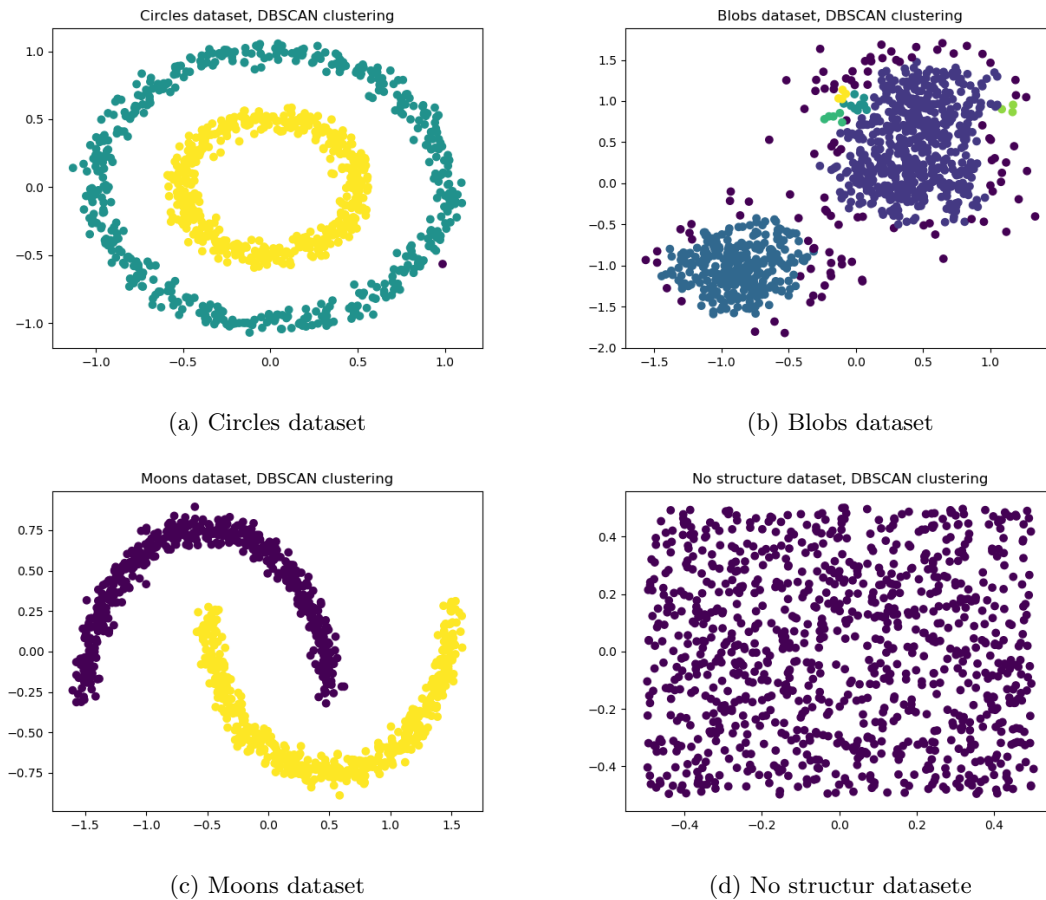


Figure 2: DBSCAN visualization

3 Research Questions

In this chapter we present the research questions, and motivate why we choose them. We also describe our goals and objectives with our thesis, and then what we thought about the expected outcome.

3.1 Research Question

RQ1: What observations can we make from the patterns, using the unsupervised clustering algorithm DBSCAN on peoples training habits with a small dataset (max 500 data points)?

RQ2: What observations can we make from the patterns, using the unsupervised clustering algorithm K-means on peoples training habits with a small dataset (max 500 data points)?

RQ3: Investigate in what fields can DBSCAN and K-means be applied?

3.2 Background

This thesis will investigate where unsupervised clustering algorithms can be applied through an literature study, this will gain the knowledge of where such algorithms can be applied. The investigation is restricted to focus on two clustering algorithms. The two algorithms use different clustering approaches, DBSCAN is density based and K-means is centroid based, which makes the investigation more interesting and also the clustering more interesting since it will be possible to get different clusters. These algorithms will cluster the dataset gathered from the survey. The limit of the dataset is set to 500 samples. This is done because we want to focus on a dataset with fewer samples. Research question 1 and 2 will provide and present an investigation through an experiment of these clustering algorithms on the dataset. The value of this is motivated by the size of the dataset. The goal with research question 1 and 2 is to see if it is possible to generalize the clusters based on a small dataset. The main objective is to investigate if we can identify patterns in the clusters. Patterns in the clusters can, for example, be if it is possible to identify what types of groups exist or to identify unknown groups with similar habits. When comparing DBSCAN with K-means we want to compare the clusters and analyze if they produce similar results.

3.3 Expectations

From the survey we expected a minimum of 100 answers. We assumed it would take a maximum of 2 weeks to receive that number of answers...

Consider research question (RQ1) we expected to observe clusters of arbitrary shapes and therefore be able to identify what groups of people have the most similar habits. find outliers in the data set, i.e. samples of data that clearly differ from the rest of the samples in the data. This is because DBSCAN have the capability to identify samples that are not close to any other samples in the dataset. We also expected to identify outliers in the dataset, i.e. samples of data that does not belong to any cluster. This is interesting because we will then be able to identify who is not belonging to any patterns and see what makes them differ from the rest.

In research question 2 (RQ2) we cluster the data with the unsupervised algorithm K-Means. We expect depending on the predefined number of clusters to give very different clusters. The reason we expect this is because the way K-means works on a dataset will put people with similar features in the same clusters but depending of the number of clustering this can be divided up to more filtered clusters.

The literature question 3 (RQ3) we read papers/articles about what fields K-means and DBSCAN can be used in. We expect to find a lot of different fields for both algorithms since both are popular. We also expect to get drastically different results from what fields the algorithms are used in because of their difference in finding cluster. The reason being that k-means is centroid based while DBSCAN is density based.

The expectation from using both algorithms is to find different generalizations. We believe that both algorithms will have no problem clustering the data, what we are unsure of is if we can get anything from looking at the clusters. The expectation we have is that K-means will make very general generalizations, this in turn could make the generalization too unspecific making the result uninteresting. To get around the result being too unspecific we will define more number of clusters to get a result that we can learn something from. DBSCAN will be less general and give us anomalies, we expect results with DBSCAN to be more specific but we are unsure if the data is too small to get specific. This could result in DBSCAN cluster either almost everything or almost nothing. To get around this we will change minpoints and epsilon to get clusters that we can see generalizations on.

4 Research Method

This thesis is carried out in two steps, one empirical part and on literature study. The first part of our thesis is the literature study, where we answer research question number 3 by reading research papers where DBSCAN and K-Means have been applied. The empirical part consists of an experiment and a survey. The purpose of the survey is to create a dataset that we can focus our thesis around. The dataset from the survey have been used during the clustering with both K-means and DBSCAN algorithms to cluster the data.

4.1 Literature study

This part describes how the literature study was conducted and its main objectives. Our goal with the literature study is to investigate in what fields and applications DBSCAN and K-means can be applied. It should be noticed that this will not cover all field they can be used. The purpose is instead to give some ideas on where it can be used and present some of this applications.

4.2 Search Engines

- Google Scholar
- BTH Summon
- Microsoft Academics

4.3 Keywords

The keywords we were using while searching for papers/articles by using search engines 4.2. The papers/articles were used to back-up our statements with other peoples statements. Our method to finding papers/articles was to type specific keywords while not making the keywords too long as we wanted to find many different fields for both algorithms.

- machine learning
- DBSCAN
- K-means economy
- K-means image segmentation
- K-means usage
- DBSCAN outlier detection
- K-means and DBSCAN

4.4 Limitations and validity threats

When determining if a literature is usable we go through the abstract, introduction and conclusion. Then we skim through the documents to see if it seems relevant enough to make a decision on using the research papers/books for the thesis. The criteria we have for our literature is that it includes nuanced to what fields DBSCAN and K-means can be used in. The literature should tell why the field can benefit from using DBSCAN and K-means.

4.5 Empirical study

4.6 Survey

The survey is the foundation for our thesis. The survey consists of 8 questions about peoples training habits. We expected between 100 - 500 answers. The survey was sent out by email to all students at BTH and also shared by social media networks such as Facebook and Slack. The survey was created with a Google Form which allowed us to obtain the results/answers as an CSV file. The survey was open for 9 days. We closed the survey when we had enough answers.

4.7 Design of survey

The most important part when designing the survey was that it would be possible to use the data in our experiment, i.e. the clustering of the data. The reason we choose training habits as our survey topic was because we thought it would increase the rate of answers, since fitness is something many people can relate to. Each question represents a feature and we decided that 8 would be enough. Both in respect to what we could expect people to answer to and to get enough number of features to get some values from the clustering. The focus of feature selection for unsupervised machine learning is to find the features that best uncovers clustering.[5] By only having relevant features for clustering we hope to achieve this.

We choose to only have one answer for each question so it would be simpler to pre-process our survey data. There is always going to be some error in making the survey[11], this can also be redundant by having a simple question. As we are working with a small dataset we both want it to be reliable. This simple approach comes with the downside that people can't pick multiple chooses if wanted to. To avoid where people would like to have more than we chooses multiple chooses such as question 4 "Mix of exercises" instead. Also we didn't want to have to many answers for each question as it could lead to much spread in data. To avoid this we picked Other in questions 1,3,4

The first part of the survey we wanted to focus on personal information question 1-3. Question 1 is there to make a big separation by having few answers. As gender is a big definer we want to follow up by spread the data more in smaller chunks by asking for age and employment.

The second part is about the training with the focus on the training. Questions 4-6,8 we focus on giving stronger identity by further spreading the data in chunks. Note in question 6 we took 0-1 hours instead of 0 hours to prevent further spread because of our assumption of having many answers whit no training habit. Question 7 was made to give each data a greater boost of identity with the training habit by only having two answers.

4.8 Experiment

This part will describe the clustering performed during our experiment and the libraries, hardware and approach. The experiment was implemented in Python3, with the machine learning library `scikit_learn` 0.19.1. This library was used because it is one of the most popular machine learning libraries REF and is it very well documented. The Python Data Analysis Library, version (pandas==0.22.0) was used for manipulating the dataset. The data used for the experiment is the dataset gathered from the survey. The dataset contained 384 samples and 8 features, with 5 of the features are categorical values and 3 features are continuous/ values. All features were used during the clustering. For our clustering we used Euclidean distance metrics. With Euclidean distance, a small distance between two objects implies a strong similarity whereas a large distance implies a low similarity. In an n-dimensional space of features the distance between two samples p and q can be calculated with:

$$dist(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2 + \dots + (p_n - q_n)^2}$$

where n is the number of features.

4.8.1 Testing environment

We used a computer with an Intel i7-7500U @ CPU 2.70GHz processor with 2 cores (4 threads) and 16GB RAM, running Linux Ubuntu 17.10 distribution. Our metrics are based on running the `scikit_learn` 0.19.1.

4.8.2 Preprocessing

This subsection describes what steps have been conducted to prepare the data for clustering. The dataset was loaded from a CSV file into a Pandas DataFrame [15], which is a two-dimensional size-mutable data structure. All data in the dataset consisted of nonnumerical data types. Categorical and non-numeric data is a problem for both K-Means and DBSCAN [16], since they work with numerical values. A way to solve the categorical features the first part of the preprocessing mapped nonnumerical values to numerical values, e.g feature Gender with the possible answers/-categories, Male = 0, Female = 1 and Other = 2. Since the algorithm will interpret Male (0) to be closer to Female (1) than it is to Other (2). The representation needs to let the computer to understand that these things are all actually equally different. The second step of the preprocessing solved this by separating the variable Gender into 3 separate variables[14] "Male", "Female" and "Other", which all can only take a binary value 0 or 1. This increased the space of dimension from 8 to 31 dimensions. Encoding categorical integer features using a one-hot (one-of-K) scheme was performed with the `scikit_learn` "`sklearn.preprocessing.OneHotEncoder`" [17] module.

The dataset contained 3 continuous features, *Age*, *How often do you workout? (days a week)*, *How much time do you spend on each workout?*. These features were mapped to a numeric representation where the lowest numerical value representing the first alternative of the respective feature and the highest numerical value was mapped to the last alternative of its respective feature.

5 Literature Review

TODO

This chapter text. Review the possible references. Write with your own words and refer to the selected references. Describe when references are supporting each other, and describe when they are in conflict with each other. Tie the references together and make your (theoretical) conclusion based on your interpretation.

6 Result

Fifth chapter text. What was the result from the empirical study, and is the theory in conflict with the results or does it support the result? Explicitly answer your Research Questions. You may divide the chapter into two separate ones, but make sure that the reader understands when you are writing about your own analysis, and when you are referring to existing references.

6.1 Literature study

This is a summary of the research papers that were found during the literature study.

TODO -Write some more to introduce this part

6.2 RQ3: In what fields can DBSCAN and K-means be applied?

A density-based algorithm for discovering clusters in large spatial databases with noise

Traffic classification using clustering algorithms[6]

This paper talks about how to accurately identify and categories network traffic according to application type. The authors describe that it is an important element of many network management tasks such as flow prioritization, traffic shaping /policing, and diagnostic monitoring. The authors describe 3 different approaches of how to identify and categorize network traffic and talks about the advantages and disadvantages of respective approach. In the first approach, the classification of network traffic is based on the mappings of known ports to applications. In the second approach, packet payload is analyzed to determine whatever they contain characteristic of known applications. The paper focusing on a third approach, were they explore how to categorize data based on only transport layer statistics with clustering. They show that cluster analysis has the ability to group and categorize network traffic using only transport layer traffic with DBSCAN and K-Means, using empirical Internet traces. The experimental results show that both K-Means and DBSCAN work very well and much more quickly then AutoClass. The results indicate that although DBSCAN has lower accuracy compared to K-Means and AutoClass, DBSCAN produces better clusters.

A New Approach of Image Segmentation Method Using K-Means and Kernel Based Subtractive Clustering Methods[4]

The paper talks about how image segmentation is the first step in image processing and a proposed machine learning algorithm to make an accurate segmentation. The researchers used clustering because of the high usage in the area, it's simplicity and efficiency. The proposed algorithm is a combination of k-means and kernel based subtractive methods. The kernel function is there to increase the efficiency by transforming the pixels from the image into other dimensions to more easily separate them. K-means is later used to identify the different types of segments 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[13]

The paper talks about with the help of machine learning are possible to find brain tumors and risk of disease. Today the findings are done using magnetic or radiation types of scans which take time and are cost heavy. If image segmentation is done by using K-means and Fuzzy C means it would both save time and make the process cheaper. This is done by crafting Brain MRI images, preprocess the image, use both algorithms and lastly predict the disease by observing the segmentation result. For finding mass tumor using K-means is enough if however there is noise in the MR image K-means needs preprocessing by filtering the noise. K-means is not good enough on it's own as it doesn't detect in detail and this is why Fuzzy C means is later used after K-means to get more accurate tumor shape extraction.

Mining smart card data for transit riders' travel patterns [?]

TODO - Literature to read

- A density-based algorithm for discovering clusters in large spatial databases with noise - DBSCANs

- <https://academic.microsoft.com/#/detail/1673310716>

- A New Approach of Image Segmentation Method Using K-Means and Kernel Based Subtractive Clustering Methods ✓

- <https://pdfs.semanticscholar.org/2ba1/942a08f3b9da97afa3b55719b5005ae2e5d0.pdf>

Note: This paper talks about image segmentation with has a lot of different areas for example brain tumor. Text talks about that K-means is powerful and efficient but not accurate enough. But if K-means is used on the correct data the results will be accurate, in this case with the help of kernel based subtractive methods.

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

- <https://pdfs.semanticscholar.org/6ff6/3b18bb5cbfc7061dcf9d956a80133c7915d0.pdf>

Note: This tells us that K-means is usable in serious fields however it doesn't do everything by itself. In this case it needs help by preprocessing and is not accurate enough to detect accurate tumor shape extraction.

- Local Search Methods for k-Means with Outliers

- <https://pdfs.semanticscholar.org/c8a7/ac02eb26260fe5559df8d2bc98b4323ba7f7.pdf>

- Mining smart card data for transit riders' travel patterns

- <https://academic.microsoft.com/#/detail/2007043321>

- Identifying and ranking the world's largest clusters of inventive activity


- http://www.wipo.int/edocs/pubdocs/en/wipo_pub_econstat_wp_34.pdf

- Understanding monthly variability in human activity spaces : a twelve-month study using mobile phone call detail records

- <https://academic.microsoft.com/#/detail/2004023904>

TODO - Write summary for each text, comparison and analysis from the literature

6.3 Survey

The survey was up between 03/07 - 03/15(9 days) 2018, we decided to close the survey from the decline from answers as seen in Figure 3. The amount of answers however was a success as previous mention we set a minimum of 100 answers for it to be passable. In the first day we had the survey shared on social media with an response of 127 answers. On the second day the mail was sent to all BTH students with 181 answers. The rest of the days are aftermath answers from the social media and mail. 

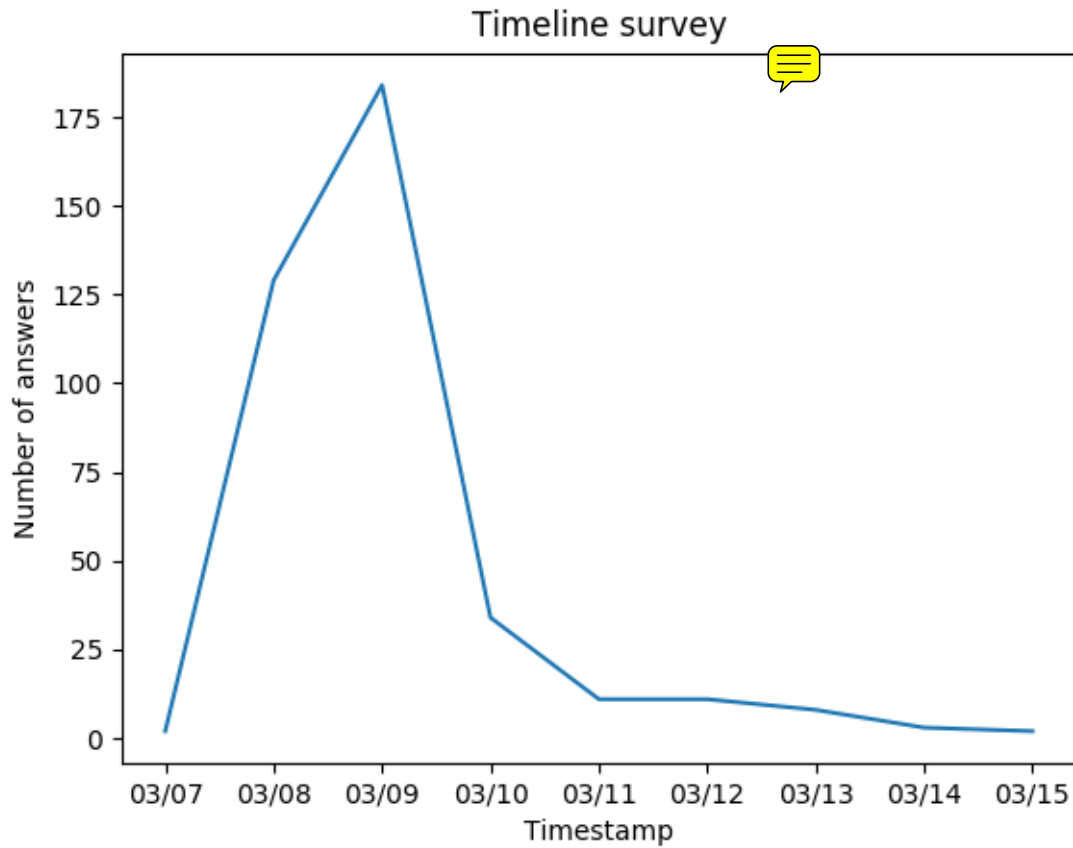


Figure 3: Timeline survey

6.4 Experiment

TODO - Test different parameters results, document all the results

7 Analysis

TODO Sixth chapter text. Be sure to make a good summary and summarize your answers to the research questions. The conclusion shall be able to be read stand alone. You should, with your conclusion, make the reader interested so that she/he read the whole thesis! The size of the conclusion part shall be $\frac{1}{2}$ - 1 page.

8 Conclusion

TODO

9 Future Work

TODO

10 Annexes

10.1 Survey Questions

In this subsection we present our survey questions. All the questions are multiple choice questions and all of them are mandatory. It is only possible to select one answer on each question.

1. Gender
 - Male
 - Female
 - Other
2. Age
 - 15-19
 - 20-24
 - 25-29
 - 30-34
 - 35-39
 - 40+
3. Employment
 - IT/Tech/Engineering
 - Construction/Mechanic
 - Economics/Business
 - Healthcare/Medicine
 - Design/Architect
 - Service/Restaurant
 - Student/Academics
 - Unemployed
 - Other
4. What do you exercise?
 - Running
 - Fitness classes (e.g. spinning, yoga, bodypump)
 - Gym
 - Crossfit
 - Mix of exercises
 - Other
 - I don't workout
5. How often do you workout? (days a week)
 - 0
 - 1-2
 - 2-3
 - 3-4

- 5+
6. How much time do you spend on each workout?
- 0-1 hours
 - 1-2 hours
 - 2-3 hours
 - 4+ hours
7. Do you workout with a partner?
- Yes
 - No
8. Why do you workout?
- Health
 - Apperance
 - Achievements
 - Enjoyment
 - Combination of above
 - None of above
 - I don't workout

10.2 Code

```
import numpy as np
import pandas as pd
from sklearn.preprocessing import OneHotEncoder, normalize
from sklearn import cluster

# Map all values to numeric data to prepare for one-hot preprocessing
def preprocess_data():
    # Get numeric data
    data = make_all_values_numeric()

    data = map_categorical_data(data)
    data = normalize_data(data)
    return data

def make_all_values_numeric():
    data = pd.read_csv("survey.txt")

    # Remove timestamp columns
    data.drop(labels=["Timestamp"], axis=1, inplace=True)

    questions = [
        ["Male", "Female", "Other"],
        ["15-19", "20-24", "25-29", "30-34", "35-39", "40+"],
        ["IT/Tech/Engineering", "Construction/Mechanic", "Economics/Business", "Gym"],
        ["Running", "Fitness classes (e.g. spinning, yoga, bodypump)", "Gym"],
        ["0", "1-2", "3-4", "5+"],
        ["0-1 hours", "1-2 hours", "2-3 hours", "4+ hours"],
        ["Yes", "No"],
        ["Health", "Apperance", "Achievements", "Enjoyment", "Combination"]
    ]

    for i, columns in enumerate(data):

        for index, element in enumerate(questions[i]):
            data[columns].replace(to_replace=element, value=index, inplace=True)

    return data

'''
Encode categorical integer features using a one-hot aka one-of-K scheme.
'''
def map_categorical_data(data):
    mask = [True, False, True, True, False, False, True, True]
    enc = OneHotEncoder(categorical_features=mask, sparse=False)
    data = enc.fit_transform(data)

    return data

def normalize_data(data):
    return normalize(data, norm="max", axis=0)

def cluster_data(data):
    # Initiate the DBSCAN model
    #dbscan = cluster.DBSCAN(eps=1.5)
    dbscan = cluster.KMeans(n_clusters=5)
    dbscan.fit(data)
```

```

    return dbscan

def get_full_information_about_outliers(data):
    temp = pd.read_csv("survey.txt")
    temp = temp.values
    for i, e in enumerate(data):
        if(e == -1):
            print(temp[i])

def main():

    # Get preprocessed data
    data = preprocess_data()

    # Cluster the data
    data_clustered = cluster_data(data)

    print("Number of samples = {}".format(len(data_clustered.labels_)))

    # Get number of outliers for DBSCAN
    outliers = [line for line in data_clustered.labels_ if line == -1]

    print(data_clustered.labels_)

    # Get number of clusters
    number_of_clusters = len(list(set(data_clustered.labels_)))

    # Print out all samples that are outliers for DBSCAN
    #get_full_information_about_outliers(data_clustered.labels_)

    print("Number of clusters: {}".format(number_of_clusters))
    print("Number of outliers = {}".format(len(outliers)))

if __name__ == "__main__":
    main()

```


References

- [1] Michael Gutmann Aapo Hyvärinen and Doris Entner. 2015.
- [2] Erik Brynjolfsson and Tom Mitchell. What can machine learning do? workforce implications. *Science*, 358(6370):1530–1534, 2017.
- [3] JOHN R BUCKLEY. Automation. *Journal of Academic Librarianship*, 20(1):40, 1994.
- [4] Nameirakpam Dhanachandra and Yambem Jina Chanu. A new approach of image segmentation method using k-means and kernel based subtractive clustering methods. *International Journal of Applied Engineering Research*, 12(20):10458–10464, 2017.
- [5] Jennifer G Dy and Carla E Brodley. Feature selection for unsupervised learning. *Journal of machine learning research*, 5(Aug):845–889, 2004.
- [6] Jeffrey Erman, Martin F. Arlitt, and Anirban Mahanti. Traffic classification using clustering algorithms. In *Proceedings of the 2006 SIGCOMM workshop on Mining network data*, pages 281–286, 2006.
- [7] 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.
- [8] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.
- [9] 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.
- [10] David Neil Lawrence Levy. *Computer Games I*. Springer, 2011.
- [11] Mark S Litwin and Arlene Fink. *How to measure survey reliability and validity*, volume 7. Sage, 1995.
- [12] 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.
- [13] Divyani Sanjay Mane and Balasaheb B Gite. Brain tumor segmentation using fuzzy c-means and k-means clustering and its area calculation and disease prediction using naive-bayes algorithm. *Brain*, 6(11), 2017.
- [14] Alexander Novikov, Mikhail Trofimov, and Ivan Oseledets. Exponential machines. *arXiv preprint arXiv:1605.03795*, 2016.
- [15] Pandas-documentation. pandas.dataframe¶. <https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.html>, March 2018.
- [16] sklearn documentation. 4.3.5. encoding categorical features¶. <http://scikit-learn.org/stable/modules/preprocessing.html#encoding-categorical-features>, March 2018.
- [17] sklearn documentation. sklearn.preprocessing.onehotencoder¶. <http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html>, March 2018.
- [18] Nainsi Soni and Manish Dubey. A review of home automation system with speech recognition and machine learning. *International Journal*, 5(4), 2017.
- [19] Hugo Steinhaus. Sur la division des corps matériels en parties. *Bull. Acad. Pol. Sci., Cl. III*, 4:801–804, 1957.