Exposé - Identifying the Ideal Length of Time to Record Smartphone Data, in Order to Obtain Distinct Clusters to Predict Eating Crises

1710601007

FH Salzburg

31st January 2020

Concept

Han, Pei, and Kamber (2011)[17, 18] declare that data mining is used to discover patterns and knowledge from data. This includes cleaning data, combining multiple sources, selecting and transforming relevant data, and extracting and evaluating data patterns.

Cluster Analysis is a type of machine learning algorithm known as unsupervised machine learning. It is used in data mining to divide data into groups (clusters). Each cluster contains data that is similar to each other, but dissimilar to the data allocated to other clusters. Cluster Analysis can be used to acquire knowledge on the distribution of the data, discover characteristics, detect outliers and reduce noise, or to pre-process data for other algorithms (Han, Pei, and Kamber 2011)[32, 362, 363, 367]. According to Bramer (2007)[311], the grouping of similar attributes is applied in various fields such as economics, marketing, medicine, crime analysis and more.

There are several different methods to create clustering. Han, Pei, and Kamber (2011)[362, 364, 366-368, 373, 374, 385, 392] explain, that objects are often arranged into clusters using distance measures (e.g. Euclidean or Manhatten distance measures). The authors divide the clustering algorithms into the following categories:

- Partitioning methods: The data is divided into k (generally pre-defined) number of groups. A data object can only be classified into one group (fuzzy partitioning methods relax this condition). Examples: k-means, k-medoids

- Hierarchical methods: Data is grouped into a hierarchy of clusters. In one approach, each object creates its own cluster and is then merged into its neighbours until all objects belong to one cluster (agglomerative or bottom-up approach). In the other approach, all objects form one cluster and are then divided until each object is contained in its own cluster (divisive or top-down approach). Examples: BIRCH, Chameleon

- Density-based methods: While partitioning and hierarchical methods only find clusters with spherical shapes, this method finds clusters with random shapes. It can also remove noise and outliers. Examples: DBSCAN, OPTICS
- Grid-based methods: The objects are quantised into grid cells. The operations are performed on the grid structure. This leads to an accelerated processing time. Examples: STING, CLIQUE

Feldman, Sanger, et al. (2007)[92] outline that the results of clustering need to be judged by a human, thus however introducing subjectivity.

SmartEater [1] is an upcoming mHealth (mobile health) app, with the goal to provide the user with content-dependent feedback, to avert a food craving episode. The app will predict future eating crises based on the user's past behaviour. In order to reduce intense user input, the app records and uses various smartphone sensor data. With the help of data mining, machine learning algorithms and pattern recognition, this recorded situational context data will aid in predicting stress. The following data is recorded by the app:

1. Background volume
2. Relative movement of the smartphone (gyro and accel)
3. Time and duration of phone calls (without storing the numbers)
4. Time of messages (e.g. SMS, WhatsApp) (without collecting identifying information such as content, addresses, numbers)
5. Screen activity (so-called touch events)
6. Screen-on-time (illuminated display)
7. Ambient brightness
8. Data volume per unit of time (summary value of all smartphone activities on the internet)
9. Switch-on and switch-off times of the smartphone

This sensor data will be recorded for different lengths of time. It is necessary to determine which time period will be most fitting to make accurate predictions for the future. This thesis will use cluster analysis to determine which time period is most significant.

According to Han, Pei, and Kamber (2011)[414], the above-mentioned clustering methods work well with data sets that are not high-dimensional an have less than 10 attributes. Since the SmartEater data set only has 9 dimensions, it is not considered high-dimensional. This paper will therefore utilise these clustering methods. Since different clustering algorithms can yield different results, multiple methods will be used and compared. To reduce the size and amount of data, dimensionality reduction will be used. Han, Pei, and Kamber (2011)[93] define

1. https://sites.google.com/site/eatingandanxietylab/resources/smarteater

dimensionality reduction as a type of data reduction, which removes random attributes and creates a smaller data set with close to equal integrity. This thesis will use principal component analysis (PCA) to reduce the dimensionality. Furthermore, T-Distributed Stochastic Neighbor Embedding (t-SNE) will be employed to depict the data set in this thesis. Maaten and Hinton (2008)[2579] first introduce t-SNE, which is used to visualise data with a higher dimensionality.

The clustering methods will be implemented using a Python machine learning platform or library (e.g. Anaconda[2], scikit-learn[3]). Next these will be implemented on the other time lengths. The resulting clusters of each time length will be compared to one another and evaluated. Evaluation examples given by Han, Pei, and Kamber (2011)[396-399] include clustering tendency, intrinsic and extrinsic measurements.

The introduction of the thesis will serve as an overview of the SmartEater project and explain how and why the subsequent experiment will be conducted. The following chapter will concentrate on the theory of data mining and cluster analysis. After covering these topics, the next section will describe the conducted experiment and its results. The conclusion will summarise the findings of the experiment.

Research Question

What is the ideal length of time to record smartphone sensor data, in order to construct distinct clusters?

Outline

1. Introduction
2. Theory
   (a) Data mining
   (b) Cluster analysis
      i. Overview of clustering algorithms
      ii. Dimensionality reduction
3. Experiment
   (a) Preparation of the data set
   (b) Clustering
   (c) Clustering after dimensionality reduction
   (d) Comparison and evaluation of clusters of different time lengths
4. Conclusions

2. `https://www.anaconda.com/`
3. `https://scikit-learn.org/stable/`

References

Bramer, Max. 2007. *Principles of data mining*. Vol. 180. Springer.

Feldman, Ronen, James Sanger, et al. 2007. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

Han, Jiawei, Jian Pei, and Micheline Kamber. 2011. *Data mining: concepts and techniques*. Elsevier.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. „Visualizing data using t-SNE.“ *Journal of machine learning research* 9 (Nov): 2579–2605.

Schedule

- 31st January 2020 - Hand in this exposé
- February 2020 - Read papers and do research
- 24th February 2020 - Upload the final exposé onto FHSys
- March 2020 - Meet with supervisor, read literature, analyse and experiment with clustering algorithms and write a rough draft
- April 2020 - Meet with supervisor, finish the paper and print and review details
- 10th May 2020 - Submission of the bachelor thesis

Supervisor

I have discussed the thesis with FH-Prof. DI Dr. Simon Ginzinger, MSc. He is working on the SmartEater research project and suggested this subject to me.