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

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. The following data is recorded by the app:

1. Background volume
2. Relative movement of the mobile phone (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

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

With the help of machine learning algorithms and pattern recognition, the recorded situational context data will aid in predicting stress.

The 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, a type of data mining, to determine which time period is most significant.

Han, Pei, and Kamber (2011)[17, 18] declares 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 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 preprocess data for other algorithms (Han, Pei, and Kamber 2011)[32, 362, 363, 367].

According to Bramer (2007)[311], grouping similar attibutes 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)[366-368, 373, 374, 385, 392, 414] divides them 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). Exmaples: k-means, k-medoids

- Hierarchical methods: Data is grouped into a hierarchy of clusters. Either each object creates its own cluster and is then merged to its neighbours until all objects belong to one cluster (agglomerative or bottom-up approach), or all objects for one cluster and are then divided until each object forms its own cluster (divisive or top-down approach). However, once a merge or split has occurred, it cannot be undone. 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

These methods work well with data sets that are not high-dimensional an have less than 10 attributes. Since the data set used in this thesis only has 9 dimensions, it is not considered high-dimensional and can be utilised in these methods.

To reduce the size and amount of data, dimensionality reduction will be used. Dimensionality reduction is 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.

Bramer (2007)[312, 313] explains, that data with a maximum number of 3 attibutes (dimensions) can easily be visualised, as can the resulting clusters. Often there is a higher number of attibutes, which is impossible to visualise.

Maaten and Hinton (2008)[2579] introduce T-Distributed Stochastic Neighbor Embedding (t-SNE) which is used to visualise high-dimensional data and will therefore be employed to depict the data set in this thesis.

In order to determine how long the smartphone sensor data should be recorded, to receive the best clustering results, the following experiment will be conducted: (— TODO - prepare the data types). Initially the data for one chosen time length will be be clustered using different clustering methods. These algorithms (— TODO - say which) will be implemented using the Python library.... (— TODO - SAY WHICH ONE). Using principal component analysis (PCA), the initial data of the same time length will be reduced and the clustering methods reimplemented. These methods will be reproduced on the other time lengths and compared using ...(— TODO - how to compare), potentially thus resulting in the time length with the clearest clusters.

The thesis will be arranged as following:

pyclustering scikit

Research Question

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

Outline

1. Introduction
2. Theory
   (a) Data Mining
   (b) Cluster Analysis
      i. Overview of Clustering Algorithms
      ii. Dimensionality Reduction
      iii. ...
3. Experiment
   (a) Preparing data
   (b) Clustering (one time length)
   (c) Clustering after Dimensionality Reduction
   (d) Comparison of clusters of different time lengths
4. Conclusions

References

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

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

FH-Prof. DI Dr. Simon Ginzinger, MSc