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 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

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 paper will use cluster analysis, a type of data mining, to determine which time period is most significant.

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.(page 17, 18) Cluster Analysis is a type of machine learning algorithm known as unsupervised machine learning. It is used to divide data into classes (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. (page 32, 362, 363, 367)

According to /textcite, clustering is used to class similar objects together, and dissimilar objects into other classes. Grouping similar attibutes is applied for various fields, such as economics, marketing, medicine, crime analysis and more. 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.

There are several different methods to create clustering. /textcite 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 high-dimensional.

In order 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 components analysis (PCA) to reduce the dimensionality.

T-Distributed Stochastic Neighbor Embedding (t-SNE) will be used to visualise high-dimensional data.

since these small dimensions - first dimensionality reduction??

Research Question

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

Outline

1. Introduction
2. Theory
   (a) UnsupervisedData Mining
   (b) Cluster Analysis
      i. Overview of Clustering Algorithms (in high dimensions ??, talk about the curse of dimensionality)
      ii. Dimensionality Reduction
      iii. ...
3. Experiment
   (a) K-Means
   (b) Hierarchical
   (c) Comparison of different lengths of time
   (d) ...
4. Conclusions

References

Albanese, Massimiliano, Angelo Chianese, Vincenzo Moscato, and Lucio Sansone. 2004. „A Formal Model for Video Shot Segmentation and its Application via Animate Vision.“ *Multimedia Tools and Applications* 24 (3): 253–272.

Bosch, Martí, Pierre Genevès, and Nabil Layaïda. 2014. „Automated refactoring for size reduction of CSS style sheets,“ 13–16. ISBN: 9781450329491. doi:10.1145/2644866.2644885.

Fried, Carrie B. 2008. „In-class laptop use and its effects on student learning.“ *Computers & Education* 50 (3): 906–914.

McConnell, Steve. 2004. *Code Complete, Second Edition.* Redmond, WA, USA: Microsoft Press. ISBN: 0735619670.

Mulloni, Alessandro, Andreas Dünser, and Dieter Schmalstieg. 2010. „Zooming Interfaces for Augmented Reality Browsers.“ In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services*, 161–170. MobileHCI '10. Lisbon, Portugal: ACM. ISBN: 978-1-60558-835-3. doi:10.1145/1851600.1851629.

Vandevoorde, David, and Nicolai M. Josuttis. 2002. *C++ Templates: The Complete Guide.* Addison-Wesley Professional, November. ISBN: 0201734842.

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