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)[18] declare that data mining is used to discover patterns and knowledge from data. 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].

There are several different methods to create clustering. Han, Pei, and Kamber (2011)[362, 364, 366-367, 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 (examples: k-means, k-medoids)
- Hierarchical methods (examples: BIRCH, Chameleon)
- Density-based methods (examples: DBSCAN, OPTICS)
- Grid-based methods (examples: STING, CLIQUE)

Bermad and Kechadi (2016) introduce in their paper, how clustering can be used in digital forensics to provide information on all the events that led up to a certain crime. They used ascending hierarchical clustering to receive clusters of events (e.g. phone calls, SMS) ordered in time, thus creating a timeline of events leading to the incident.

Dey and Chakraborty (2015)[1,2,6,7] give an example, where clustering was implemented to predict future weather. Air pollutant data was preprocessed and then arranged into clusters using (incremental) DBSCAN clustering. Finally, priority based protocol was used on them to

predict weather conditions and a temperature range. The accuracy of the technique, based on hit and miss times, was calculated to approximately 74.5%.

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 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

1. https://sites.google.com/site/eatingandanxietylab/resources/smarteater
2. https://www.anaconda.com/
3. https://scikit-learn.org/stable/

lengths. The resulting clusters of each time length will be compared to one another and evaluated. Berkhin (2006)[39] states, that the Silhouette Coefficient (Kaufman and Rousseeuw 2009)[87] can be used to measure the separation between clusters.

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. Secondly, existing work relating to this subject will briefly be presented. 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. Related work
3. Theory
   (a) Data mining
   (b) Cluster analysis
      i. Overview of clustering algorithms
      ii. Dimensionality reduction
4. Experiment
   (a) Preparation of the data set
   (b) Clustering
   (c) Clustering after dimensionality reduction
   (d) Comparison and evaluation of clusters of different time lengths
5. Conclusions

References

Aranganayagi, S., and K. Thangavel. 2007. „Clustering Categorical Data Using Silhouette Coefficient as a Relocating Measure.“ In *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*, 2:13–17. December. doi:10.1109/ICCIMA.2007.328.

Berkhin, P. 2006. „A Survey of Clustering Data Mining Techniques.“ In *Grouping Multidimensional Data: Recent Advances in Clustering*, edited by Jacob Kogan, Charles Nicholas, and Marc Teboulle, 25–71. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-540-28349-2. doi:10.1007/3-540-28349-8_2. https://doi.org/10.1007/3-540-28349-8_2.

Bermad, N., and M. T. Kechadi. 2016. „Evidence analysis to basis of clustering: Approach based on mobile forensic investigation.“ In *2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, 300–307. December. doi:10.1109/SETIT.2016.7939884.

Bublitz, C. F., A. C. Ribeiro-Teixeira, T. A. Pianoschi, J. Rochol, and C. B. Both. 2017. „Unsupervised Segmentation and Classification of Snoring Events for Mobile Health.“ In *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 1–6. December. doi:10.1109/GLOCOM.2017.8255031.

Chen, Y., J. Kim, and H. S. Mahmassani. 2014. „Pattern recognition using clustering algorithm for scenario definition in traffic simulation-based decision support systems.“ In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 798–803. October. doi:10.1109/ITSC.2014.6957787.

Cohen, Michael B., Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. 2015. „Dimensionality Reduction for K-Means Clustering and Low Rank Approximation.“ In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, 163–172. STOC '15. Portland, Oregon, USA: Association for Computing Machinery. ISBN: 9781450335362. doi:10.1145/2746539.2746569. https://doi.org/10.1145/2746539.2746569.

Dey, R., and S. Chakraborty. 2015. „Convex-hull DBSCAN clustering to predict future weather.“ In *2015 International Conference and Workshop on Computing and Communication (IEMCON)*, 1–8. October. doi:10.1109/IEMCON.2015.7344438.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. „A density-based algorithm for discovering clusters in large spatial databases with noise.“ In *Kdd*, 96:226–231. 34.

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

Karimzadeh, M., Z. Zhao, F. Gerber, and T. Braun. 2018. „Mobile Users Location Prediction with Complex Behavior Understanding.“ In *2018 IEEE 43rd Conference on Local Computer Networks (LCN)*, 323–326. October. doi:10.1109/LCN.2018.8638045.

Kaufman, Leonard, and Peter J Rousseeuw. 2009. *Finding groups in data: an introduction to cluster analysis.* Vol. 344. Hoboken, New Jersey: John Wiley & Sons.

Lu, E. H., and V. S. Tseng. 2009. „Mining Cluster-Based Mobile Sequential Patterns in Location-Based Service Environments.“ In *2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, 273–278. May. doi:10.1109/MDM.2009.40.

Lu, E. H., V. S. Tseng, and P. S. Yu. 2011. „Mining Cluster-Based Temporal Mobile Sequential Patterns in Location-Based Service Environments.“ *IEEE Transactions on Knowledge and Data Engineering* 23, no. 6 (June): 914–927. ISSN: 2326-3865. doi:10.1109/TKDE.2010.155.

Luna-Romera, José María, María del Mar Martínez-Ballesteros, Jorge García-Gutiérrez, and José C. Riquelme-Santos. 2016. „An Approach to Silhouette and Dunn Clustering Indices Applied to Big Data in Spark.“ In *Advances in Artificial Intelligence*, edited by Oscar Luaces, José A. Gámez, Edurne Barrenechea, Alicia Troncoso, Mikel Galar, Héctor Quintián, and Emilio Corchado, 160–169. Cham: Springer International Publishing. ISBN: 978-3-319-44636-3.

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

Mahmud, M. S., M. M. Rahman, and M. N. Akhtar. 2012. „Improvement of K-means clustering algorithm with better initial centroids based on weighted average.“ In *2012 7th International Conference on Electrical and Computer Engineering*, 647–650. December. doi:10.1109/ICECE.2012.6471633.

Patel, KM Archana, and Prateek Thakral. 2016. „The best clustering algorithms in data mining.“ In *2016 International Conference on Communication and Signal Processing (ICCSP)*, 2042–2046. IEEE.

Patel, Vaishali R, and Rupa G Mehta. 2011. „Impact of outlier removal and normalization approach in modified k-means clustering algorithm.“ *International Journal of Computer Science Issues (IJCSI)* 8 (5): 331.

Sajana, T, CM Sheela Rani, and KV Narayana. 2016. „A survey on clustering techniques for big data mining.“ *Indian journal of Science and Technology* 9 (3): 1–12.

Sapkota, N., A. Alsadoon, P. W. C. Prasad, A. Elchouemi, and A. K. Singh. 2019. „Data Summarization Using Clustering and Classification: Spectral Clustering Combined with k-Means Using NFPH.“ In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 146–151. February. doi:10.1109/COMITCon.2019.8862218.

Wang, Bangjun, Li Zhang, Caili Wu, Fan-zhang Li, and Zhao Zhang. 2017. „Spectral clustering based on similarity and dissimilarity criterion.“ *Pattern Analysis and Applications* 20, no. 2 (May): 495–506. ISSN: 1433-755X. doi:10.1007/s10044-015-0515-x. https://doi.org/10.1007/s10044-015-0515-x.

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.