FH Salzburg
MultiMediaTechnology

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

Bachelor Thesis 2

Author: Natasha Lauren Troth
Advisor: FH-Prof. DI Dr. Simon Ginzinger, MSc.

Salzburg, Austria, 10.05.2020

Affidavit

I herewith declare on oath that I wrote the present thesis without the help of third persons and without using any other sources and means listed herein; I further declare that I observed the guidelines for scientific work in the quotation of all unprinted sources, printed literature and phrases and concepts taken either word for word or according to meaning from the Internet and that I referenced all sources accordingly.

This thesis has not been submitted as an exam paper of identical or similar form, either in Austria or abroad and corresponds to the paper graded by the assessors.

____
Date

____
Signature

____
First Name *Last Name*

Kurzfassung

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean venenatis nulla vestibulum dignissim molestie. Quisque tristique tortor vitae condimentum egestas. Donec vitae odio et quam porta iaculis ut non metus. Sed fermentum mauris non viverra pretium. Nullam id facilisis purus, et aliquet sapien. Pellentesque eros ex, faucibus non finibus a, pellentesque eu nibh. Aenean odio lacus, fermentum eu leo in, dapibus varius dolor. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Proin sit amet ornare velit. Donec sit amet odio eu leo viverra blandit. Ut feugiat justo eget sapien porttitor, sit amet venenatis lacus auctor. Curabitur interdum ligula nec metus sollicitudin vestibulum. Fusce placerat augue eu orci maximus, id interdum tortor efficitur.

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean venenatis nulla vestibulum dignissim molestie. Quisque tristique tortor vitae condimentum egestas. Donec vitae odio et quam porta iaculis ut non metus. Sed fermentum mauris non viverra pretium. Nullam id facilisis purus, et aliquet sapien. Pellentesque eros ex, faucibus non finibus a, pellentesque eu nibh. Aenean odio lacus, fermentum eu leo in, dapibus varius dolor. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Proin sit amet ornare velit. Donec sit amet odio eu leo viverra blandit. Ut feugiat justo eget sapien porttitor, sit amet venenatis lacus auctor. Curabitur interdum ligula nec metus sollicitudin vestibulum. Fusce placerat augue eu orci maximus, id interdum tortor efficitur.

Contents

List of Figures

Listings

List of Tables

1 Introduction

Han, Pei, and Kamber (2011)[18, 32, 362, 363, 367] 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.

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).

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 up 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

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

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 establish 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 and 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 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 thesis will be structured as follows: The first section will briefly present existing work relating to this subject. 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. In the final sections, the findings of the experiment will be discussed and summarised.

!!!WRITE ABOUT EATING DISORDERS.., ALSO WRITE ABOUT MOBILE HEALTH APPS

2 Related work

Related Work

page 3 book from libraary

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

3 Theory

in theory

3.1 Data mining

Larose and Larose (2015)[4] declares that data mining is used to recognise patterns and trends in large amounts of data.

Data mining requires continuous human supervision for quality monitoring and evaluation. Data mining software alone will server wrong results.

Data mining is used for description of patterns and trends, estimation of numerical values, prediction of future results, classification of categorical variables, clustering of similar objects and association of attributes.

There are two types of data mining methods: supervised and unsupervised. The majority of methods are supervised. In supervised methods, there is a predefined target variable. The method receives several examples, where the target variable value is defined, thus learning which values of the target variable correspond to which values of the predictor variable. The goal of the unsupervised approach is to find patterns and structure in the inserted variables. Therefore, no target variable is established. Clustering is the most known unsupervised method. Problems that can occur in data mining methods are data dredging and overfitting. Data dredging is when false results arise in data mining due to random variations of data. Cross-validation is used to prevent data dredging, by guaranteeing that the results can be generalised to an independent data set. Overfitting arises, when the provisional model trys to fit perfectly to the training model, thus leading to the accuracy being higher on the training set than on the test set. ??? BIAS-VARIANCE TRADE-OFF

3.1.1 Data preprocessing

Data sets first need to undergo a data preprocessing step, including data cleaning and data transformation. This aids in making the data useful in data mining. Raw data extracted directly from databases can be incomplete (values are missing) or be noisy (contains outliers), or may contain out-dated or redundant data. This unpreprocessed data may also not be in a correct form for data mining models. The goal is to decrease garbage in, garbage out (GIGO). Reducing the irrelevant data that is fed into the model (garbage in), the amount of irrelevant data received out of the model is reduced (garbage out).

There are some data mining that have trouble functioning correctly when fed outliers. Moreover, outliers may be data errors. Graphical methods used to identify outliers include, histograms or two-dimensional scatter plots. The Z-score method can be used to numerically calculate outliers. Outliers should not automatically be removed from the data set.

Data cleaning is used to handle outliers, errors and unusual values found in the data set.

One approach to handle records with missing values, is to delete said record. The author does not recommend this, since it could lead to a biased subset of data, if the missing values are systematic. Furthermore, it would mean wasting the data stored in the other fields of that record. A preferred approach is to substitute the missing value. The value can be replaced, either with a constant determined by the data analyst, with a field mean (for numerical values) or mode (for categorical values), with a random value, or with imputed values based on different features of the record. Replacing missing values can be a gamble, since it can possibly lead to invalid results. For example, the authors experimented with a database of cars. Substituting a missing brand with a random value (here ”Japan”) led to a car, that doesn't even exist. Data imputation takes into account the other attributes stored in the record and from these, calculates what the missing value would most likely be.

Another step in data preprocessing is identifying misclassifications. An example given by the authors, is classifying a record as USA instead of US, or France instead of Europe. These classes only contained one record in comparison to the other more frequently used classes.

In some data mining algorithms, variables with higher ranges can unjustly influence the results, having more influence than smaller ones. Therefore, the authors recommend to normalise numerical data.

Han, Pei, and Kamber (2011)[105-106] describes normalisation as giving the attributes equal weight. For example, it can transform the data to fall in a smaller, common range (e.g. [-1, 1]). It therefore hinders variables with large ranges from outweighing ones with smaller ranges. For example, income would have a larger range than binary attributes. Typical normalisation techniques include *min-max normalization*, *Z-score standardization* and *decimal scaling*. For the following examples, A is a numerical attribute from a data set, a single value of this attribute is represented with v_i:

- Min-max normalization uses linear transformation to normalise the original value to a newly defined minimum ($newMin_A$) and maximum ($newMax_A$) value (e.g. 0.0 and 1.0). The minimum and maximum value found in A are presented as min_A and max_A:

  $$v_i' = \frac{v_i - min_A}{max_A - min_A}(newMax_A - newMin_A) + newMin_A$$

  If new data is added, that isn't within the min and max of A range, an ”out-of-bounds” error will occur.

- Z-score (or zero-mean) normalization normalises the values using the mean ($\bar{A}$) and standard deviation σ_A of A.

  $$v_i' = \frac{v_i - \overline{A}}{\sigma_A}$$

  The advantage of this normalisation method, is that the min and max of A do not need to be known, or when there are outliers that would overrule the min-max method.

- The decimal scaling method moves the decimal point as many spaces, so that the maximum absolute attribute value of A is below zero. The smallest number of digits that the decimal point has to be moved, so that the largest absolute number in A is below zero, is represented by j:

$$v_i' = \frac{v_i}{10^j}$$

According to Larose and Larose (2015)[39-41, 45], flag variables can be used to transform categorical variables into numerical. A flag variable can take on one of two values: 0 and 1 (e.g. female = 0, male = 1). When k¿=3 (k being the amount of categorical predictors), the variables can be transformed into k-1 flag variables. Assigning categorical variables numerical values is not advised, since this orders the categorical variables. For example, if North = 1, East = 2, South = 3 and West = 4, West would be closer to South than to North, etc.

ID fields should be removed from the dataset, since the value is different for each record and not helpful.

Han, Pei, and Kamber (2011)[16, 17] explain, that the term ”data mining” is a misnomer. A more suitable phrase would be ”knowledge mining from data”. The word ”mining” represents valuable nuggets found within large amounts of raw material. Other names used to describe the same process include: knowledge discovery from data (KDD), knowledge extraction, data/pattern analysis, data archaeology, and data dredging. According to the authors, the discovery of data is an iterative process represented in the following steps

1. Data cleaning
2. Data integration (combine multiple data sources)
3. Data selection (relevant data is extracted)
4. Data transformation (into applicable forms for data mining)
5. Data mining (discover patterns)
6. Pattern evaluation (determine if patterns have a meaning)
7. Knowledge presentation

Typical data forms used for mining can be database data, data warehouse data, and transactional data. Other forms include data streams, ordered/sequence data, graph or networked data, spatial data, text data, multimedia data, and the World Wide Web.

Outliers are objects that vary to the general behaviour or model of the data. In some cases, the uncommon events are of more interest. One of these instances is detecting unusually large payments compared to the card holders normal payments, to uncover fraudulent usage of credit cards.

Using unsupervised machine learning, it is possible to detect classes within data.

3.2 Dimensionality reduction

According to Larose and Larose (2015)[92, 93], Bellman Bellman (2015) explains, that the data in high dimensional spaces is sparse. .....FIND SOMEWHERE THAT THIS MEANS THE CURSE OF DIMENSIONALITY..... High dimensional data sets arise, when a database has multiple variables. A high amount of predictor variables in a model can make the interpretation of an analysis more complicated. It can lead to overfitting and overlooking crucial relationships between predictors. Furthermore, visualising higher dimensions becomes more challenging. High quality visualisation methods usually cannot depict more than five dimensions. Humans use these visualisations for visual pattern recognition. Dimensionality reduction techniques have the ability to reduce the number of predictor items, aid in ensuring that these predictor items are independent, and present a framework for interpretability of the results.

Principal components analysis (PCA) is a dimensionality reduction method.

As stated by Han, Pei, and Kamber (2011)[93, 95-96], dimensionality reduction is a data reduction method. Data reduction is utilised to attain a smaller, more concentrated data set, whilst mostly keeping the integrity of the initial data set. PCA projects the initial data onto a smaller space, thus removing random variables. The data it is applied to can be ordered or unordered, sparse and skewed. PCA is conducted in the following steps:

1. The first step is to standardise the input data, therefore making the data-range identical. Larose and Larose (2015)[94] declares, that after standardising the data, the mean is zero and the standard deviation is one.

2. Next, k orthonormal vectors are calculated, the so called *principal components*. These unit vectors present a basis for the input data, which are a linear combination of the principal components. Larose and Larose (2015)[94] explain, that the principal components can be discovered, by rotating the initial coordinate system to the direction of maximum variability. These then create a new coordinate system.

3. In the following step, as stated by Han, Pei, and Kamber (2011)[95-96], the principal components are put into order by their decreasing significance/strength, thus presenting their variance. These vectors are used as new axes for the data, the first axis exhibits the highest variance.

4. Due to the decreasing order of variance, the vectors with the lowest variance can be removed, therefore reducing the amount of data and number of dimensions. Despite the loss of data, the components with higher variance can approximate the original data.

Other methods of dimensionality reduction include wavelet transforms (e.g. discrete wavelet transform (DWT)).

3.3 Cluster Analysis

in cluster analysis

3.3.1 Overview of clustering algorithms

Overview Clustering Algs

4 Experiment

in experiment

4.1 Preparation of the data set

in prep of data set

4.2 Clustering

in clustering

4.3 Clustering after dimensionality reduction

in clustering after dim red

4.4 Comparison and evaluation of clusters of different time lengths

in comparison of diff time lengths

5 Discussion

in discussion.tex

6 Conclusion

in conclusion

References

Bellman, Richard E. 2015. *Adaptive control processes: a guided tour.* Princeton university press.

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. Hammamet, Tunisia, December. doi:`10.1109/SETIT.2016.7939884`.

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. Vancouver, BC, Canada, October. doi:`10.1109/IEMCON.2015.7344438`.

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

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

Larose, Daniel T, and Chantal D Larose. 2015. *Data mining and predictive analytics.* 2. ed.. Wiley series on methods and applications in data mining. Hoboken, New Jersey: John Wiley & Sons. ISBN: 9781118116197.

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

Appendices

Anhänge löschen, die nicht verwendet werden.

A git-Repository

Das Repository dient zur Dokumentation und Nachvollziehbarkeit der Arbeitsschritte. Stellen Sie sicher, dass der/die BetreuerIn Zugriff auf das Repository hat. Stellen im Sinne des Datenschutzes sicher, dass das Repository nicht für andere zugänglich ist.

Verpflichtende Daten für Bachelorarbeit 1 und 2:

- LaTeX-Code der finalen Version der Arbeit
- alle Publikationen, die als pdf verfügbar sind.
- alle Webseiten als pdf

Verpflichtende Daten für Bachelorarbeit 2:

- Quellcode für praktischen Teil
- Vorlagen für Studienmaterial (Fragebögen, Einverständniserklärung, ...)
- eingescanntes, ausgefülltes Studienmaterial (Fragebögen, Einverständniserklärung, ...)
- Rohdaten und aufbereitete Daten der Evaluierungen (Log-Daten, Tabellen, Graphen, Scripts, ...)

Link zum Repository auf dem MMT-git-Server `gitlab.mediacube.at`:

`https://gitlab.mediacube.at/fhs123456/Abschlussarbeiten-Max-Muster`

B Vorlagen für Studienmaterial

Vorlagen für Studienmaterial müssen in den Anhang.

C Archivierte Webseiten

`http://web.archive.org/web/20160526143921/http://www.gamedev.net/page/resources/_/technical/game-programming/understanding-component-entity-systems-r3013`, letzter Zugriff 1.1.2016

`http://web.archive.org/web/20160526144551/http://scottbilas.com/files/2002/gdc_san_jose/game_objects_slides_with_notes.pdf`, letzter Zugriff 1.1.2016