Susanne Jauhiainen

Knowledge Discovery from Physical Activity

Master's Thesis in Information Technology

May 31, 2017

University of Jyväskylä

Department of Mathematical Information Technology

Author: Susanne Jauhiainen

Contact information: susanne.m.jauhiainen@student.jyu.fi

Supervisor: Tommi Kärkkäinen and Sami Äyrämö

Title: Knowledge Discovery from Physical Activity

Työn nimi: Knowledge Discovery from Physical Activity

Project: Master's Thesis

Study line: Laskennalliset tieteet

Page count: 99+0

Abstract: In this master's thesis the Knowledge Discovery in Databases (KDD) process and its usage with physical activity data are discussed. The KDD process has multiple steps, including preprocessing, transformation, and data mining. Clustering is used as the data mining technique and is introduced in detail. A large set of different Cluster Validation Indices (CVAIs) and their implementations are tested with the k-means clustering and the best performing ones further generalized. In the empirical part, physical activity data from Finnish seventh-grade students is assessed following the KDD process and using multiple different transformations with different clustering methods. The aim is to find out, whether unsupervised data mining can help detect novel and useful information from this data.

Keywords: Knowledge discovery, physical activity, cluster validation index

Suomenkielinen tiivistelmä: Tässä pro gradu -tutkielmassa käydään läpi Knowledge Discovery in Databases (KDD) -prosessi ja sen soveltamismahdollisuuksia fyysiseen aktiivisuuteen liittyvän datan kanssa. KDD-prosessi koostuu monesta eri vaiheesta, sisältäen esikäsittelyn, datan muunnoksen ja tiedonlouhinnan. Tässä tutkielmassa tiedonlouhinnan menetelmänä käytetään klusterointia, joka käydään läpi yksityiskohtaisesti. Vertailemme myös laajan joukon eri klusterointi-indeksejä (CVAIs) sekä niiden eri toteutuksia k-means klusteroinnin kanssa ja esittelemme parhaat näistä yleisemmässä muodossa. Tutkielman empiirisessä osassa seitsemäsluokkalaisten koululaisten aktiivisuusdataa tutkitaan KDD-prosessia seuraten

ja hyödyntäen monia eri datan muunnoksia ja klusterointimenetelmiä. Tarkoituksena on selvittää, voiko ohjaamattoman tiedonlouhinnan avulla löytää uutta ja hyödyllistä informaatiota datasta.

Avainsanat: Knowledge discovery, fyysinen aktiivisuus, klusterointi-indeksi

Glossary

KDD	Knowledge Discovery from Databases
EDA	Exploratory Data Analysis
LDA	Linear Discriminant Analysis
SVM	Support Vector Machine
CPM	Counts Per Minute
CVAI	Cluster Validation Index
IoT	Internet of Things
QS	Quantified Self
PCA	Principal Component Analysis
MET	Metabolic Equivalent of Task
MVPA	Moderate to Vigorous Physical Activity

List of Figures

List of Tables

Contents

1 Introduction

Physical activity can be defined as "any bodily movement produced by skeletal muscles that requires energy expenditure" (Caspersen, Powell, and Christenson 1985). The amount and form of daily physical activity, over the whole life span, can greatly affect individuals health and quality of life, which makes the assessment of it very important. Active lifestyle can, for example, reduce the risk of coronary heart disease, hypertension, diabetes, some cancers, and premature mortality in general (Health and Services 1996). According to the World Health Organization (WHO), physical inactivity is the fourth biggest global risk for mortality, responsible for 5.5% - over 3 million - of all deaths (WHO 2009).

As the technical abilities to measure activity's different components and recognize its different forms have advanced, many applications and studies in a variety of domains have emerged. For example in gerontology, monitoring the activity of elderly people can be used to detect if they fall (Mubashir, Shao, and Seed 2013) or if they have chest pain or headache (Khan and Sohn 2011). This nondiscruptive monitoring can be done from outside their free-living conditions, which improves the quality of their lifes and saves elderly care resources by supporting home care. In sports the foot-ground contact time and the stride ratecadence can be analysed when running and utilized in maximizing the performance (Morin et al. 2007; Weyand et al. 2000). Heart rate monitoring has also become very general and can give a lot of information about, for example, the intensity of physical activity (Vesterinen 2016).

There are many different components of human activity that can be measured with a variety of devices such as accelerometers or heart rate monitors. For example, the total physical activity, the duration, frequency and intensity of physical activity, energy expenditure, as well as number of steps, speed and distance when walking (Butte, Ekelund, and Westerterp 2012) can be determined from the raw sensory measurements. Furthermore, the locomotive activities (e.g. walking, jogging, running) can be classified into their own categories (Bao and Intille 2004; Kwapisz, Weiss, and Moore 2011).

Measuring of these components of human activity can be done objectively with many sensors which can be divided into external and wearable ones (Lara and Labrador 2013). Firstly,

inertial sensors are wearable sensors based on inertia, with accelerometers being the most widely used example (Avci et al. 2010). Acceleration data can be used to estimate, e.g., the intensity of physical activity over time (Chen and Bassett 2005).

Secondly, sensors for physiological signals can be used in measuring the activity. These are normally wearable sensors. Heart rate monitors, for instance, have developed rapidly in recent years and are today very common in sports and training. They are mainly used to determine the exercise intensity (Achten and Jeukendrup 2003), but heart rate variability can also be used to analyze how well an athlete is adapting to endurance training (Vesterinen et al. 2013) or in prevention and detection of overtraining (Achten and Jeukendrup 2003). Thirdly, vision sensors are external sensors used, for example, in gait analysis and other machine vision applications (Poppe 2007).

With the increase of different human activity sensors and monitors, massive streams of data about our lives are being created continously. This has led to suggestion of the concept Internet of Humans (Arbia et al. 2015), in connection with the Internet of Things (IoT). While IoT is the general term referring to the things and devices being connected to the internet, IoH can be thought as humans being connected to internet by using different monitors and loading data about themselves to the internet. This personal data can also be referred to as MyData (Poikola and Honko 2010). An emerging trend is to talk about so called quantified-self (QS), which has been definded "as any individual engaged in the self-tracking of any kind of biological, physical, behavioral, or environmental information" (Swan 2013). As the amount of data grows, new approaches are needed to utilize it more effectively.

Knowledge Discovery in Databases (KDD) is a process of finding valid and novel information from large amounts of data (Fayyad, Piatetsky-Shapiro, and Smyth 1996b). It has many steps, beginning from preprocessing the data, transforming it and finally by using data mining techniques in order to find patterns that can be interpreted into knowledge. Compared to the more traditional analysis techniques, data mining is rather about finding novel and unexpected information from the data than about confirmation of some existing hypotheses (Hand, Mannila, and Smyth 2001). KDD's potential to utilize the large datasets, that are now common in the field of physical activity, makes it a natural framework for the knowledge discovery from such sources.

As people become more inactive and the sedentary time increases (Owen et al. 2010), possible dangers of sedentariness have been recognized lately (Helajärvi et al. 2013; Rezende et al. 2014). Sedentary time can be defined as activities with a very low energy expenditure (e.g., 1.0–1.8 metabolic equivalents (METs) (Jans, Proper, and Hildebrandt 2007)) or as time spent sitting/supine (Chastin and Granat 2010). The physical activity and sedentary behaviour of children has become a concern (Hillman, Kamijo, and Scudder 2011) and a very current research topic. According to a Finnish study (Husu, Vähäpyä, and Vasankari 2016), 7-13 year-old children spend more than half of their waking hours sedentary, mainly sitting. It is important to recognize the reasons behind and ways to decrease the sedentariness. Clustering has the potential to bring out conjunctive factors in groups, that might be related to the sedentary behaviour of the students in that group.

In this study, the activity behaviour of Finnish seventh-grade students is being assessed. The aim is to, by following the KDD process and using clustering as the data mining technique, find out different activity profiles amongst the students. These profiles are constructed based on both the activity and sedentary behaviour separately. Therefore, the research question is as follows:

- **Can we find novel and useful information from students activity data using unsupervised data mining?**

For finding the number of clusters present in the data, we tested and further developed many existing Cluster Validation Indeces (CVAIs), as well as developed our own index (Jauhiainen and Kärkkäinen 2017). Our further generalization of the CVAIs enables their easy usage with many distance measures.

The structure of this thesis is as follows: in Chapter 2, the KDD process and its steps are introduced and in Chapter 3, clustering in more detail. Ways for measuring and quantifying physical activity through a variety of sensors and some example applications in the field of physical activity are presented in Chapter 4. The empirical part of the thesis is stated in Chapters 5 and 6, evaluation of the CVAIs for clustering the students activity behaviour and the schoolchildren's activity study, respectively. Finally, conclusions and some discussion about potential future work are given in Chapter 7.

2 The KDD process

Knowledge discovery in databases (KDD) was originated in the 1990s, after the increase of large, ubiquitous databases (Piatetsky-Shapiro 1990). As this growing amount of data was of very little value in its raw form and becoming unmanagable for the existing data analysis techniques, new methods for knowledge extraction were well needed. The original KDD process was introduced by Fayyad et. al (Fayyad, Piatetsky-Shapiro, and Smyth 1996a; Fayyad, Piatetsky-Shapiro, and Smyth 1996b; Fayyad, Piatetsky-Shapiro, Smyth, et al. 1996) and defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". So data itself is not knowledge, but after identifying these patterns, information, from data, they can be interpreted into knowledge.

Even thought data mining and knowledge discovery are nowadays often used as synonyms (Piatetsky-Shapiro 2000), in this thesis the KDD process is used to refer to the overall process of discovering useful knowledge, while data mining is a specific step in the process where an algorithm is applied to the data for pattern discovery (Piatetsky-Shapiro 1990).

The KDD process is iterative and interactive, having multiple steps with, as stated above, data mining being one of them. Before the actual mining, data selecting, preprosessing and transforming can be done, while afterwards interpretation of patterns is required for gaining knowledge. The steps of KDD process are outlined in Figure 1.

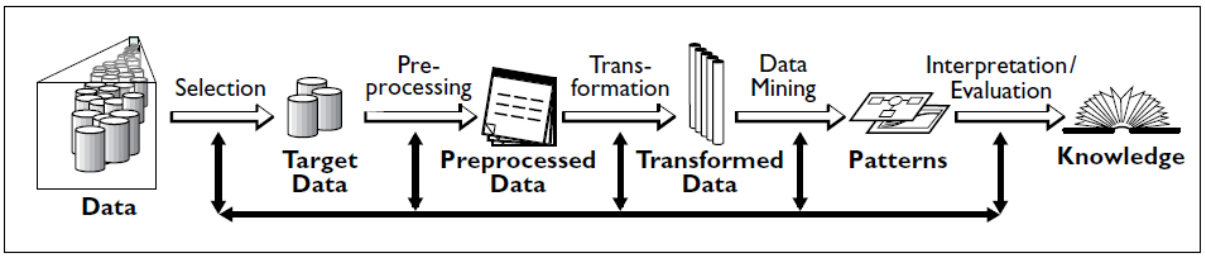

Figure 1. The original KDD-process. From Fayyad, Piatetsky-Shapiro, and Smyth 1996b.

2.1 Data matrix

Data is most often represented in the form of a matrix, where observations are as row vectors and their attributes on the columns. Let the i:th observation be $\mathbf{x}_i = \{x_{i,j}\}$, where its attributes $j = 1,\ldots,n$, for $i = 1,\ldots,N$. So data with N observations having n attributes could be represented in an $N \times n$ matrix

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,j} & \ldots & x_{1,n} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,j} & \ldots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i,1} & x_{i,2} & \ldots & x_{i,j} & \ldots & x_{i,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \ldots & x_{N,j} & \ldots & x_{N,n} \end{bmatrix} \in \mathbb{R}^{N \times n}.$$

The rows, observations, can also be referred as points, objects, or feature vectors. The attributes of the observation in turn can be called as features, variables, or dimensions. The number of columns, n, is considered as the dimensionality of the data whereas the number of observations, N, is considered as the amount of the data. Throughout this work we denote a vector with the notation $\mathbf{x}$ and a matrix with $\mathbf{X}$.

The attributes can be of many different forms, of which one usually distinguish the following types (Bramer 2007; Zaki and Meira Jr 2014):

- Categorical
  - Nominal: Categories with no order. These can be represented also in numerical form, but have no mathematical interpretation (e.g., blue/red/green).
  - Ordinal: Similar to nominal, but with an order (e.g., child/adult/elderly).
  - Binary: Special case of the nominal variable, with only two classes represented, for example, by 0 or 1, true or false (e.g., pregnant/not pregnant).
- Continuous
  - Integer: Variables that are integers and have arithmetic meaning (e.g., number of students).

  - Interval-scaled: Numerical values with equal intervals from zero or other origin point (e.g., temperature in Celsius).
  - Ratio-scaled: Similar to interval-scaled but zero represents the absence of measurement (e.g., temperature in Kelvin).

2.2 Preprocessing

First step in the knowledge discovery process is the selection of the target datasets, according to the goal and utilizing background information about the application domain (Fayyad, Piatetsky-Shapiro, and Smyth 1996b). After the target set has been chosen some data cleaning is often needed, as raw data can be noisy and messy. For example, missing values, corrupted values, or improper sampling can exist.

As nowadays' large, often sparse datasets have a lot of missing values due to variety of reasons, the decision on how to handle these missing points is important. Missing points can occur as a result of human errors in measuring, but most often the values just are unavailable (e.g., sensor has not been worn during shower or one has no answer to a certain question (Dixon 1979)). These missing data values can be divided into three categories (Roderick JA Little and Rubin 2014):

- Missing completely at random (MCAR) – the missingness does not depend on the data values, either missing or observed
- Missing at random (MAR) – missingness depends only on the data components that are observed, not on those that are missing
- Not missing at random (NMAR) – missingness depends on the missing values in the data.

The missing points can be handled, for example, by replacing them using imputations based on either statistics (e.g. mean, median) or predictive modelling (Batista and Monard 2003). Some other options are to omit the whole observation including missing values or just to use data as it is with techniques that are able to handle these missing points (Roderick JA Little and Rubin 1989).

The data can also include incorrect, corrupted, and even impossible datapoints, that can be caused either by human errors or by sensors giving erroneous measurements (Hammer 1976). These points could include very large values, outliers, that violate common sense (e.g., age 167 years, temperature 140 celcius degrees), or impossible combinations (e.g. gender: male, pregnant: yes). These datapoints can be replaced similarly as missing values or set to missing values (Äyrämö 2006). Whichever strategy for missing and spurious variable handling is chosen, the impacts on the further process should also be considered carefully (Bramer 2007).

Data filtering is an important step in the preprocessing, but it can mean different things and be done very differently in different application areas. For example, in web usage mining it can be the identification and removal of robots' requests (Tanasa and Trousse 2004), or with textual data stopwords (Nieminen, Pölönen, and Sipola 2013) can be filtered out. When considering accelerometer data there is almost always some high frequency noise and the data itself consists of two components, gravitational and body acceleration (Yang, Wang, and Chen 2008). The high frequency noise can be removed, for example, using median (Karantonis et al. 2006) or Gaussian (Luo and Hu 2004) filtering. In addition, the components of gravitational and body acceleration can be separated using high pass (Yang, Wang, and Chen 2008), low pass (Karantonis et al. 2006), or band pass filtering.

2.3 Transformation

Data transformation can include, for instance, scaling, normalization, data projection, or dimension reduction. It can also include the forming of different transformed representations of the data. These representations can be made by summing or otherwise conducting new variables or by arranging and combining the data in novel ways (Han, Pei, and Kamber 2011). Summing can be done, for example, over time summing observation per second to observations per minute etc. All these transformed representations can enable the finding of new, unexpected structures and patterns from data, compared to the raw form (Hand, Mannila, and Smyth 2001).

Manipulation of the distribution of variables might be an useful transformation. For exam-

ple, taking a logarithmic transformation of the data is very common for skewed data (Hand, Mannila, and Smyth 2001). This makes the distribution more symmetric, smoothing differences between large and small values and so the larger ones do not dominate in data mining. For more evenly distributed data (i.e. not skewed) the distribution can be transformed by normalizing it to follow the standard normal distribution:

$$x'_{i,j} = \frac{x_{i,j} - \bar{x}_j}{\sigma_j}, \tag{2.1}$$

where $i = 1\ldots,N$, $j = 1\ldots,n$, and $\bar{x}_j$ is the mean and σ_j the standard deviation of the jth column.

Range normalization can also be done by min-max scaling (Zaki and Meira Jr 2014), which is particularly handy when dealing with distances. The data can be scaled into the range between $[0,1]$ by

$$x'_{i,j} = \frac{x_{i,j} - mn}{mx - mn}, \tag{2.2}$$

where $i = 1,\ldots,N$, $j = 1,\ldots,n$, and mn and mx are the minimum and maximum values of the jth column:

$$mn = min(\{(\mathbf{x}_i)_j\}_{i=1}^N), \quad mx = max(\{(\mathbf{x}_i)_j\}_{i=1}^N). \tag{2.3}$$

In general, the normalization of data between the range $[a,b]$ can be done with the following

$$x'_{i,j} = (b-a)\frac{x_{i,j} - mn}{mx - mn} + a. \tag{2.4}$$

2.3.1 Dimension reduction

The importance of dimension reduction originates from the information overload that is a result of advances in data collection and storage capabilities during past decades (Verleysen

and François 2005). This increase in the amount of data is caused both by samples collected over time and the number of attributes that can be measured and stored.

The purpose of dimension reduction is to reduce the number of variables under consideration by attaining a set of principal variables (Fayyad, Piatetsky-Shapiro, and Smyth 1996b), which then makes further processing easier. Dimension reduction can be divided into feature selection, which aims to finding a subset of original variables that best represents the original data, and feature extraction (Liu and Motoda 1998).

Feature selection is a technique where a subset of the original features is selected for further processing (Hand, Mannila, and Smyth 2001) so that the selected subset represents the data as accurately as possible. The most simple feature selection technique is to choose features based on intuition and domain expertise, but also many methods for the selection have been developed. For example, LASSO (Friedman, Hastie, and Tibshirani 2001) is a regression analysis method that can be used in variable selection. Another technique, based on integrating the derivative of the feedforward mapping with respect to inputs over the training data, was introduced in (Kärkkäinen 2015).

Principal component analysis (PCA) is a common feature extraction based dimension reduction teqhnique invented in the beginning of 1900s (Pearson 1901). The main goal of PCA is to find a subset of attributes, principal components, from the data so that the most relevant information can still be represented with it and no important information is lost. It tries to find these components so that they cover as much as possible of the overall variance of the data.

Following the notation in (Kärkkäinen and Saarela 2015), as we have the original data, a set of N vectors $\{\mathbf{x}_i\}$ in $\mathbb{R}^n$, the aim is to transfer these to a new set, $\{\mathbf{y}_i\}$ in $\mathbb{R}^m$, so that $m < n$. We are looking for a set of orthonormal basis vectors $[\mathbf{u}_1\mathbf{u}_2\ldots\mathbf{u}_n]$ and with $z_k = \mathbf{u}_k^T\mathbf{x}$, we can denote $\mathbf{x} = \sum_{k=1}^{n} z_k\mathbf{u}_k$.

Considering a new vector $\tilde{\mathbf{x}} = \sum_{k=1}^{m} z_k\mathbf{u}_k + \sum_{k=m+1}^{n} b_k\mathbf{u}_k$, where the last term is the residual error $\mathbf{x} - \tilde{\mathbf{x}} = \sum_{k=m+1}^{n}(z_k - b_k)\mathbf{u}_k$, we have the least-squares-error (LSE):

$$\frac{1}{2}\sum_{i=1}^{N}\|\mathbf{x_i}-\tilde{\mathbf{x}}_i\|^2=\frac{1}{2}\sum_{i=1}^{N}\sum_{k=m+1}^{n}(z_{i,k}-b_k)^2=\frac{1}{2}\sum_{k=m+1}^{n}\mathbf{u}_k^T\mathbf{C}\mathbf{u_k}. \tag{2.5}$$

Here $\mathbf{C}$ is the sample covariance matrix of the data

$$\mathbf{C}=\sum_{i=1}^{N}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^T, \tag{2.6}$$

where $\bar{\mathbf{x}}$ is the mean vector. Let $\{\lambda_k,\mathbf{u}_k\}$ be the kth eigenvalue and eigenvector of the symmetric matrix $\mathbf{C}$. Hence, the eigenvalues and eigenvectors satisfy the following eigenvalue problem

$$\mathbf{C}\mathbf{u}_k=\lambda_k\mathbf{u}_k,\qquad k=1,\ldots,n. \tag{2.7}$$

Utilizing the equation 2.5 and the orthogonality of $\mathbf{u}_k$s we can write the LSE of equation 2.3 as

$$\frac{1}{2}\sum_{k=m+1}^{n}\lambda_k. \tag{2.8}$$

So, the m eigenvectors that correspond to the m largest eigenvalues of $\mathbf{C}$ form the basis for the transformed representation. Then we have the transformed data points as $\mathbf{y}_i=\mathbf{u}_k(\mathbf{x}_i-\bar{\mathbf{x}})$, where $i=1,\ldots,m$ and $\mathbf{u}_1$ is the basis vector corresponding to the largest eigenvalue, $\mathbf{u}_2$ to the second largest and so on.

2.4 Data mining

Data mining is the technical step of the KDD process where patterns are being discovered from the preprocessed and transformed data by applying some specific algorithms. It has been defined as "the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" (Hand, Mannila, and Smyth 2001). So, while more traditional

data analysis techniques often aim at confirmation of predefined hypotheses, data mining is more about finding novel, unexpected information from the data. Also, as massive databases, with data of various forms and qualities, grow in number, these traditional techniques might become insufficient.

Data mining techniques can be either predictive or descriptive. The purpose of a descriptive model is to summarize and describe the whole data (Hand, Mannila, and Smyth 2001). This can be done, for example, with statistics, using mean, median, mode, or standard deviation. Descriptive techniques focus on understanding the underlying features and processes in data to get insight on how to approach the future.

Predictive techniques, in turn, use the data to make predictions about unknown future events. With data, the value of a particular variable (output) can be predicted from the values of the known variables (input). The nature of predictive modeling is probabilistic and it often uses statistical techniques. So the purpose is not to predict what will happen in the future but rather what might happen. Usually both descriptive and predictive techniques are used together in data mining applications.

Data mining can also be divided into supervised and unsupervised (Bramer 2007). In supervised learning the aim is to use labelled data to predict a value for unseen observation, while in unsupervised learning the labels are inferned from the data itself (Bramer 2007). Examples of supervised learning are classification and regression while clustering and exploratory data analysis are unsupervised.

According to the KDD process by Fayyad et. al (Fayyad, Piatetsky-Shapiro, and Smyth 1996a; Fayyad, Piatetsky-Shapiro, and Smyth 1996b), data mining algorithms are a specific mix of the following three components:

- **Model.** Discoverable patterns containing parameters determined from the data.

- **Preference criterion.** Criteria for how well a particular pattern meets the goal of the KDD process. Basis for preference of one model or set of parameters over another. Can be based, for instance, on the accuracy, novelty, utility, or understandability of the model.

- **Search algorithm** Definition of an algorithm for (i) parameter search and (ii) model search. After the specification of the model and preference criterion, the job of the search algorithm is purely an optimization task.

Next the data mining tasks adapting Zaki (Zaki and Meira Jr 2014) are introduced, with clustering in more detail in its own chapter.

2.4.1 Exploratory data analysis

In exploratory data analysis (EDA) the purpose is to explore the data to find interesting and unexpected structures from it, often using statistical methods (Hand, Mannila, and Smyth 2001). It was first promoted in 1970's, encouraging statisticians to explore the data and formulate hypotheses that could spring new data collections and experiments (Tukey 1980). So, as statistical methods are often used for the confirmation of predefined hypoteses, EDA tries to analyze the data so that new hypotheses can be suggested based on it. It is an approach for analyzing data to summarize the main characteristics in it and the detection of structures is most often done using visualization methods such as box plots (see Fig. 2), histograms, different charts and scatter plots.

With high-dimensional data the visualization has to be done for few dimensions at a time or do dimension reduction first. In fact, as data mining often deals with excessively high-dimensional data sets, another goal of exploratory analysis can be to reduce the amount of data.

The numeric attributes in data can be analyzed with basic statistical methods. These include univariate, bivariate, and multivariate analysis, considering one, two, or more attributes at a time. In the univariate analysis, measures of central tendency (sample mean, expected value, median, mode) and measures of dispersion (range, variance and standard deviation, variance of the sample mean) are calculated and analyzed. In the bivariate analysis, as the focus is on two attributes at the same time, the association and dependence between them is of special interest. Measures of location and dispersion (mean, variance) and measures of associations (covariance, correlation) are used. In the multivariate analysis, similar measures as in the

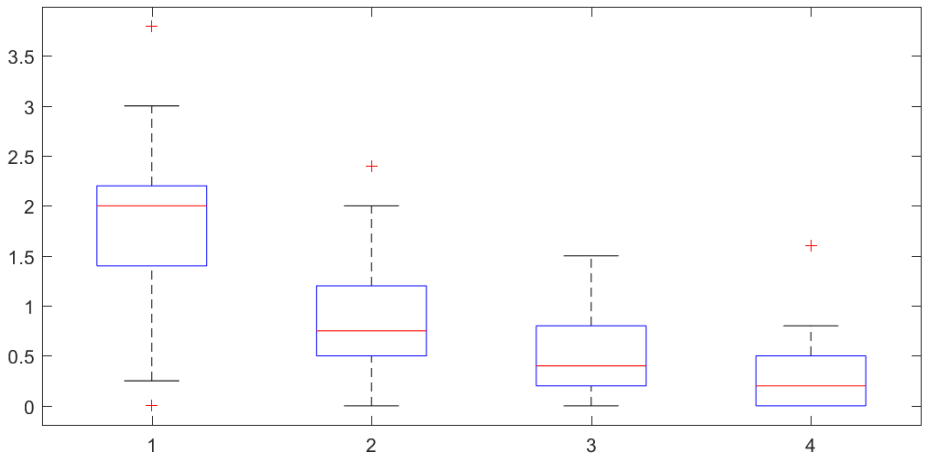

Figure 2. With box plot one can get a quick overview of the data. The upper and lower bounds of the box are the third and first quartile, the line inside the box is the median, the whiskers represent max and min values, and the plus signs stand for outliers. This box plot is derived from the sedentary data of students in Chapter 6.

bivariate analysis are used, but with more attributes at a time (Zaki and Meira Jr 2014). With bivariate and multivariate analysis data normalization is often necessary, especially if the values are different in scale. Data can be normalized by scaling it so that all values are inside a certain range, e.g., $[-1, 1]$ or by manipulating it to be normally distributed (see Section 2.3).

2.4.2 Frequent pattern mining

Frequent pattern mining is a task where informative, interesting, and useful patterns are extracted from large and complex datasets (Zaki and Meira Jr 2014). A pattern can be interesting if it, for example, appears frequently, or in turn is more rare but with higher confidence. The main goal is to find hidden and novel trends and behaviours from the data to understand it better.

There are two main types of patterns that can be discovered from a database, frequent itemsets and sequential rules (Bramer 2007). Frequent itemsets consist of co-occurring attributes and a common example of frequent itemset mining is market basket analysis (Aggarwal and

Han 2014). Market basket analysis tries to find out what items are often purchased together and after mining and analyzing these itemsets, *associations rules* can be extracted. These rules can then be exploited by the shop owners, for example, by placing the items bought together close to each other in the store.

Adapting the notation in (Goethals 2003), let $I = \{i_1, i_2, ..., i_n\}$ be an itemset containing n items and $D = \{t_1, t_2, ..., t_m\}$ a database with m transactions, where transactions contain a subset of items from I. A *rule* is defined as $X \rightarrow Y$, where $X, Y \subset I$ and $X \cap Y = \emptyset$. In the itemset mining, these rules are extracted from the databases with the help of constraint values, called *support (sup)* and *confidence (conf)*. Here the support value indicates how often a certain item or itemset occurs in a database and the confidence value, defined as $conf(X \rightarrow Y) = sup(X \cup Y)/sup(X)$, indicates how often a certain rule is fulfilled

Table 1. An example database with five transactions containing five items.

Transaction ID	Milk	Bread	Coffee	Beer	Apples
1	1	1	0	0	1
2	0	0	1	0	1
3	1	0	1	1	0
4	0	0	1	0	0
5	0	1	0	1	0

To demonstrate this with an example, a sample database has been defined in Table 1. In the market basket analysis a rule can be, for example, that "if people buy milk and bread, they also buy apples", marked as $\{Milk, Bread\} \Rightarrow \{Apples\}$. So here the itemset $\{Milk, Bread\}$ occurs in one of the five transactions and the support value is therefore $1/5 = 0.2$. Considering a rule $\{Milk, Bread\} \Rightarrow \{Apples\}$ the confidence is defined as $\sup\{Milk, Bread\}/\sup\{Apples\}$ so $\text{conf}(\{Milk, Bread\} \Rightarrow \{Apples\}) = 0.2/0.2 = 1$. This means that in all the cases in the database that a customer bought milk and bread, he also bought apples and the extracted rule is that if someone buys milk and bread they will also buy apples with 100% certainty.

In mining, a minimum support value is given by the user and if the support of certain itemset

is greater than the minimum support, then this itemset is frequent. The most used algorithm for itemset mining is Apriori (Agrawal, Srikant, et al. 1994). Apriori algorithm starts with identifying frequent individual items in the database and proceeds by extending the item into a set of items that often occur together with the individual item.

Sequential rule mining in turn is about discovering frequent subsequences, from a sequence database (Mabroukeh and Ezeife 2010). In sequential rules the attributes in a sequence have some relationships, for instance, temporal or positional, with each others. It is used in many real-world applications, such as, text mining (sequence of letters) or bioinformatics (DNA or protein sequences). There are many methods for sequence minining of which some allow gaps between the elements of a sequence and some do not.

Frequent sequence mining can be demonstrated by string mining. String mining deals with a limited number of characters or symbols and a sequence or a string is defined as an ordered list of these characters. A sequence containining k characters is often called k-sequence and a substring is a part of sequence, having less than k characters. With a database of N sequences the support of a certain subsequence is defined as the total number of sequences containing this subsequence. In an example database containing seven sequences, $\{milk, carrot, bread, avocado, apples, cake, beer\}$, the subsequence 'ca' is found in three of them and hence the support $sup(ca) = 3$. The relative support in turn is the fraction of sequences containing a subsequence, in this case $rsup(ca) = 3/7$. Again, if the support is greater than an user-defined minimum, the subsequence is discovered as frequent.

2.4.3 Classification

Classification is a task of predicting the class of a given point based on known points (Tan and Steinbach 2006). It is a supervised data mining method as the classes for classification are given a priori. A classical example of a classification problem is the Fisher's Iris data (Fisher 1936), including 150 flowers/observations and having their sepal length, sepal width, petal length, and petal width as variables. There are three classes, Iris setosa, Iris versicolor, and Iris virginica, and the classes for all 150 flower are known. The classes can be distinguished with a linear discriminant model based on the combination of the four features.

There exists a variety of different classification methods, such as probabilistic classifiers, decision trees, linear discriminant analysis, support vector machines, and so on (Zaki and Meira Jr 2014). Of probabilistic classifiers the naive Bayes (Bramer 2007) and nearest neighbors (Hand, Mannila, and Smyth 2001) classifiers are the most well known. The key idea of these classifiers is to assess the class of a certain point based on the classes of close observations. With naive Bayes the closeness is measured by probability based on the Bayes' theorem and with nearest neighbor it is measured with distance. So if the most of the observations close to the point belong to a certain class, it is probable that this point also belongs to the same class. A simple example data is given in Fig. 3, having two classes that are quite separable. The black circle in the figure would be classified into class one and the square into class two, based on the classes of near observations.

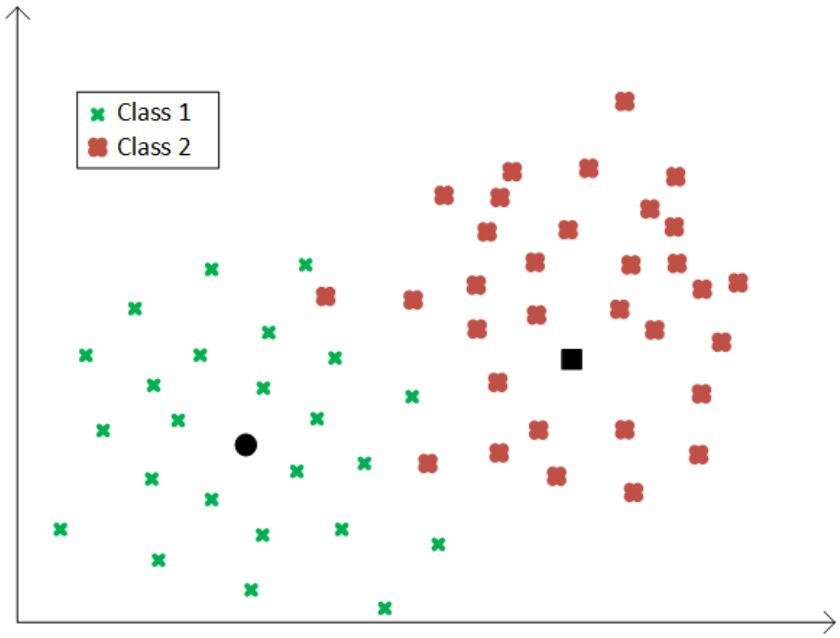

Figure 3. Data with two classes, black circle and square are to be classified.

Decision trees use tree-like graphs consisting of decisions and their possible consequences (Han, Pei, and Kamber 2011). In classification the leaves represent class labels and the branches represent rules that lead to a certain class label (Rokach and Maimon 2014). For example, the classification of unknown fruits can be seen in Fig. 4. Note that the decision tree can also be probabilistic if the tree provides the posterior class probability distribution at each note (Hand, Mannila, and Smyth 2001).

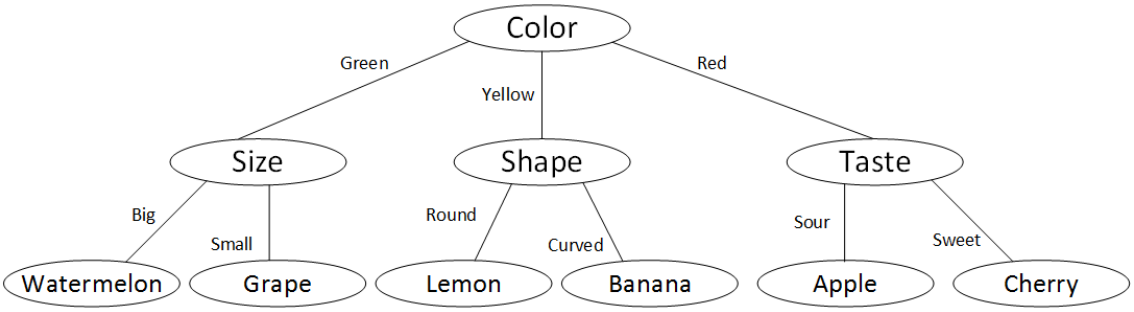

Figure 4. Fruit classification using decision tree

In linear discriminant analysis (LDA) the goal is to find a vector w that maximizes the separation between classes when projected onto w (Xanthopoulos, Pardalos, and Trafalis 2013), while in support vector machines (SVM) we try to find a hyperplane that maximizes the separation, or margin, of classes (Cortes and Vapnik 1995) (see Fig. 5). While other classifiers consider all data points, SVM focuses on the points that are closest to the separating plane (called support vectors) and so the most difficult to tell apart.

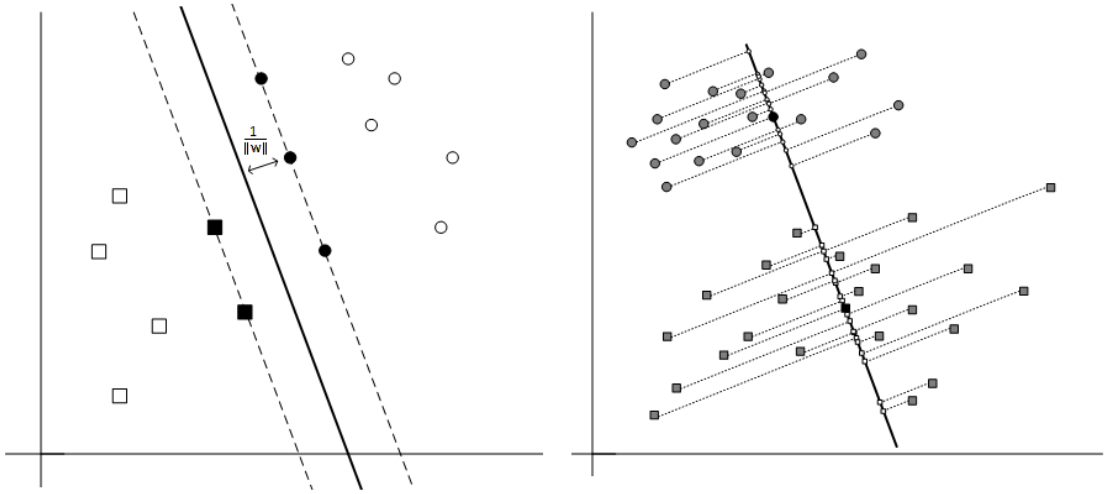

Figure 5. On the right LDA with vector w and on the left SVM, where grey markers are the support vectors and $\frac{1}{\|w\|}$ is the margin. Adapted from (Zaki and Meira Jr 2014)

2.4.4 Regression

Regression analysis is a statistical process for estimating the relationships between variables in data (Montgomery, Peck, and Vining 2015). It is similar to classification but, while in classification the aim is to predict a discrete class label, in regression the aim is to predict a continuous value based on the data (Tan and Steinbach 2006).

In simple linear regression the goal is to fit a straight line to the data (Montgomery, Peck, and Vining 2015). The equation for this is

$$y = \beta_0 + \beta_1 x, \tag{2.9}$$

where βs are unknown coefficients, x denotes the independent variables and y the dependent variables whose values are predicted. As an example, the delivery time of a product can be predicted based on the delivery volume (see Figure 6).

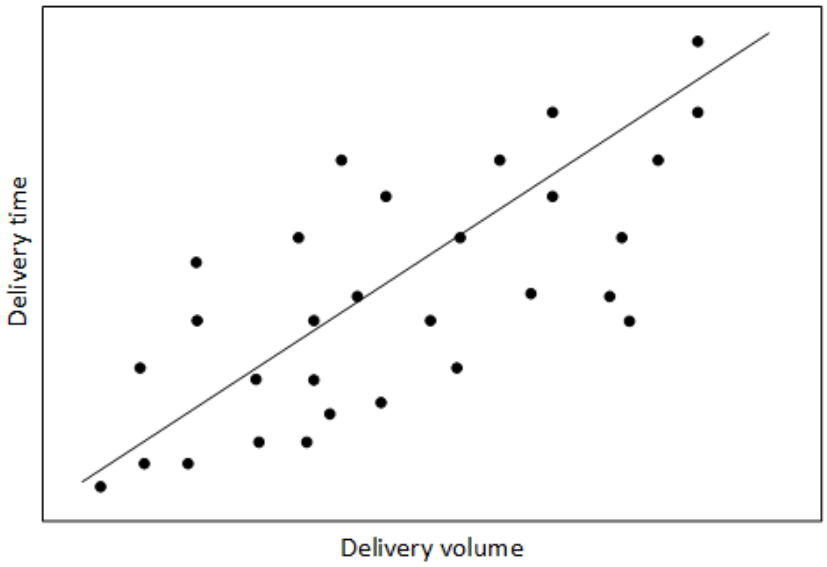

Figure 6. A line fitted to describe the relationship between delivery time and delivery volume. Adapted from Montgomery, Peck, and Vining 2015

2.5 Summary

In this chapter, the KDD process and its steps were introduced. KDD is "the nontrivial process of indentifying valid, novel, potentially useful, and ultimately undestandable patterns in data" (Fayyad, Piatetsky-Shapiro, and Smyth 1996b). It consists of multiple steps, including data mining where the actual discovering of patterns and new information from data is done. Before data mining the data needs to be preprocessed to handle missing and spurious data values and filter out any noise. Often also transformations are done and the transformed form of data can enable the finding of new interesting patterns compared to the raw form.

Data mining is the step in KDD where the patterns and information are being discovered from the preprocessed and transformed data. Data mining technique can refer, for example, to exploratory data analysis, frequent pattern mining, classification or clustering, which is introduced in the next chapter. The data mining technique should be chosen case dependently, as well as the preprosessing and transformation that is done.

3 Clustering

Clustering is unsupervised classification of observations, data items, or feature vectors into groups (Jain, Murty, and Flynn 1999). These groups are called clusters and constructed by the clustering algorithm during the procedure. The purpose is to form these clusters so that "patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster" (Jain, Murty, and Flynn 1999). There exists many measures for this similarity as the concept of 'cluster' cannot be precisely defined (Estivill-Castro 2002) and, as a consequence, also variety of clustering algorithms exist.

3.1 What is a cluster

A common problem is the desicion of how many clusters there are in the data and as said in (Estivill-Castro 2002), the "clusters are, in large part, on the eye of the beholder". As demonstrated in Fig. 7, the amount of clusters is not always clear and many 'right' answers exist. In the figure, with Gaussian distribution the number of clusters could be claimed to be for instance three or seven and with the Laplacian distribution even four or six because there is more noise making the case ambiguous. In this case, the number of clusters can also depend on the resolution, i.e., are the similarities within and between clusters considered locally or globally.

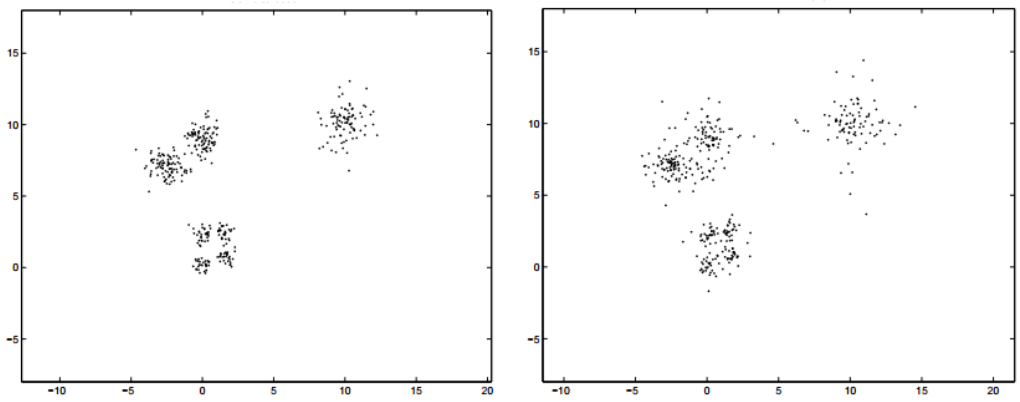

Figure 7. On the left, data drawn from Gaussian distribution and on the right from Laplacian distribution. From Äyrämö 2006, pp. 56

The most common measure for similarity is distance, suitable with cases where the observations in cluster are close to each other, as in Fig. 7. Distance itself can also be defined with different measures, and the p-norm of a vector $\mathbf{x}$ is defined as (adapted from Kärkkäinen and Heikkola 2004)

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{N} |\mathbf{x}_i|^p \right)^{1/p}. \tag{3.1}$$

Now, if $p = 1$ we have the so-called cityblock distance and for $p = 2$ the euclidean distance. Consider the following optimization problem

$$min J_p^q(x), \quad where \quad J_p^q(x) = \sum_{i=1}^{N} \|\mathbf{x} - \mathbf{x}_i\|_p^q. \tag{3.2}$$

Now, if both $p = q = 2$, we end up with the data mean, if $p = q = 1$ the median, and if $p = 2q = 2$ the spatial median. The data mean always has a unique value while the median value is unique for odd N (Kärkkäinen and Heikkola 2004) but not for even N. For spatial median, the existence and uniqueness in the case of non-collinear data is proved in (Äyrämö 2006, Theorem 4.6.1 and 4.6.2).

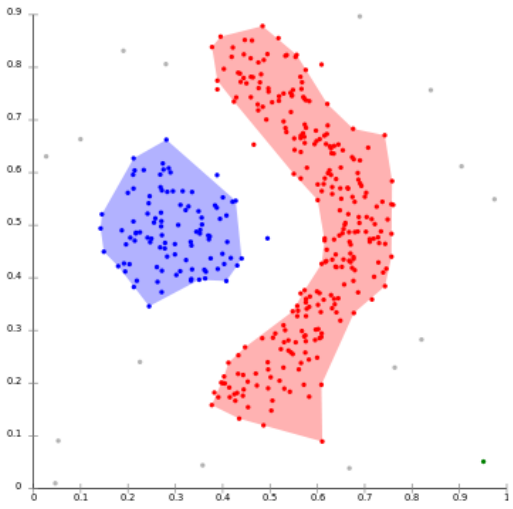

Figure 8. Density-based dataset with two non-linearly separable clusters. From https://en.wikipedia.org/wiki/DBSCAN.

The distance measure as well as the measure for similarity has to be chosen case dependently, based on the nature of the data to be clustered. For instance, another popular measure for similarity is density, suitable in cases such as in Fig. 8, where the points in same cluster are of similar, higher, density.

3.2 Hierarchical clustering methods

Hierarchical clustering creates a sequence of nested partitions that can be visualized by a tree or a dendrogram (Zaki and Meira Jr 2014). These methods try to find a hierarchy of clusters either with an agglomerative or divisive strategy (Jain and Dubes 1988). In the agglomerative approach, each data point is separate at the beginning and by merging these together the clusters are formed. The merging is done to the two closest clusters until all points are members of the same cluster or if specified, when there are exactly k clusters remaining. The number of clusters is decreased by one in every step, resulting in a sequence of nested clusterings. Divisive methods in turn start with one cluster and perform recursive splitting for the increased number of clusters.

3.3 Density-based clustering methods

In density-based datasets, for example in Fig. 8, the clusters are not of linear shapes and two points from different clusters can have smaller distance than two points in the same cluster. The areas with higher density are considered as clusters whereas the more sparse areas are considered as border points and noise (Kriegel et al. 2011).

The most widely used density-based clustering method is Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al. 1996). With DBSCAN a data point is defined as a core point, if it has at least a $minpts$ number of neighbors within the distance of ε, so in its so called ε-neighborhood. Those points that do not meet the $minpts$ treshold, but belong to the ε-neighborhood of some core point, are defined as border points. Points that are neither core or border points are considered as noise or outliers. The ε-neighborhood of a point x can be defined as follows:

$$N_\varepsilon(x) = \{y \mid d(x,y) \leq \varepsilon\}, \tag{3.3}$$

where $d(x,y)$ is the distance between x and y. A simple example of the determination of points can be seen in Fig. 9.

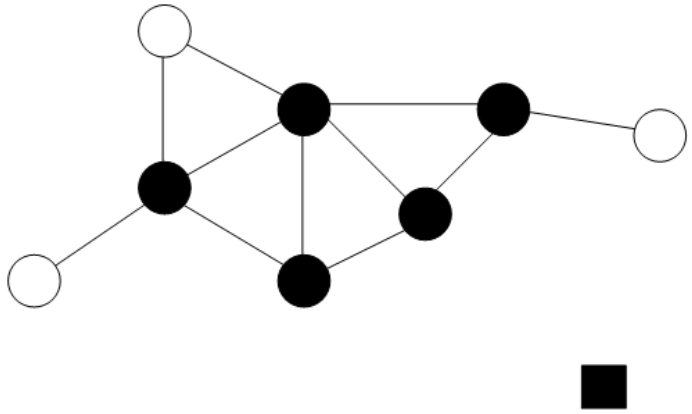

Figure 9. The DBSCAN algorithm when $minpts = 3$, where the black circles are core points, white ones are border points and the black square is noise. The lines implicate a distance less than or equal to ε.

3.4 Prototype-based clustering methods

The goal of prototype-based clustering methods is to partition the data directly into given amount of clusters, which are represented by the prototypes (Zaki and Meira Jr 2014). When dividing the data into k clusters, the process can be outlined as follows (see, e.g., Aldenderfer and Blashfield 1984):

1. Initialize k cluster prototypes
2. Assign each observation in data into closest of the k prototypes
3. Recompute the prototypes
4. Repeat steps 2 and 3 as long as the prototypes change or an user-defined maximum number of iterations is reached

The repeated steps 2 and 3 are done so that they minimize the within-cluster error, also referred to as clustering error. So, in other words they resolve the following (Friedman,

Hastie, and Tibshirani 2001):

$$\underset{\{b_k\}_{k=1}^{K}}{\arg\min}\ J(\{\mathbf{b}_k\}), \tag{3.4}$$

where

$$J(\{\mathbf{b}_k\}) = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in \mathbf{C}_k} \|\mathbf{x}_i - \mathbf{b}_k\|_p^q = \sum_{k=1}^{K} J_k. \tag{3.5}$$

K is the number of clusters, $\mathbf{x}_j$ is an observation assigned to cluster k and (i.e., $\mathbf{b}_k$ is the closest prototype). J_k is the clustering error of cluster k and J is the sum of these, so the total clustering error being minimized. Also, let $\{\mathbf{c}_k\} = \{\mathbf{b}_k^*\}$ be the local minimizer of (3.4) and $J(\{\mathbf{c}_k\}) = J^*$ the local minimum with J_k^* the within-cluster final error.

For the whole data, let *argmin* $J(b_0)$, where $J(b_0) = \sum_{i=1}^{N} \|\mathbf{x}_i - b_0\|_p^q$, be the corresponding problem and thus $c_0 = b_o^*$ the global minimizer and $J(c_0) = J_0^*$ the global minimum for this problem.

In this study partitional clustering methods, namely k-means (MacQueen et al. 1967) and k-medians, have been used. K-means is a very popular and simple partitional method that aims to dividing the data into k clusters such that each observation belongs to the cluster with nearest mean. It is obtained from Equation 3.5 by choosing $p = 2$ and $q = 2$. The k-medians method in turn follows from selections $p = 1$ and $q = 1$ and k-spatialmedians from $p = 2$ and $q = 1$ (Äyrämö 2006) (see Section 3.1).

3.4.1 Cluster initialization

In prototype-based clustering, the initial placement of cluster prototypes has a significant effect on the clustering as with different initializations, different results can be obtained (Celebi, Kingravi, and Vela 2013) and there is no guarantee of a converge to a global minimum of the clustering error (Celebi, Kingravi, and Vela 2013; Jain 2010).

Many methods for cluster initialization have been developed, but a general strategy is to

use random initialization with several runs (Xu and Wunsch 2005). However, the way of computing the initial prototypes should be chosen case-dependently (Saarela and Kärkkäinen 2015), so that the solution will be good also globally. For example, choosing the initial prototypes as far away from each other as possible might result into a good solution with respect to the between-cluster-error, i.e., the members in a cluster will be dissimilar to the members in other clusters(Arthur and Vassilvitskii 2007).

Because the initialization effects the obtained results so much and the converge is only local, multiple repetitions are used. The clustering method is being repeated with different initialization and the best, i.e., smallest, value for (3.4) is then chosen as the global minimum.

3.5 Robust clustering

In statistics, robustness means that a technique has a good performance with data from a wide range of different distributions. Robust techniques are insensitive to small deviations in the assumptions (Huber 1981). With modern large datasets the distributions are often not normal and the data can contain a lot of missing values and outliers. As normal methods and measures are unfavorably affected by these, robust methods provide an alternative.

For example, with the k-means method, the prototype of a cluster is represented by the cluster mean and mean can be affected a lot by outliers. Therefore, more robust and reliable prototype-based methods are sometimes needed, such as k-medians. In general a robust prototype-based method can be obtained from Equation 3.5 by choosing $q = 1$ (Äyrämö 2006). Moreover, a straightforward approach referred as *available data strategy*, introducing no extra assumptions to deal with missing values, was proposed and thoroughly tested in (Roderick J Little and Rubin 1987; Äyrämö 2006).

3.6 Cluster validation

As clustering is an unsupervised data mining technique, with no predefined classes, the results have to be somehow validated. The validation task is to find the partition that fits best the nature of data. Cluster validation indices (CVAIs) are measures used for determining the

number of clusters in the data. These indeces can be approached based either on external or internal criteria (Halkidi, Batistakis, and Vazirgiannis 2002). The external criteria is based on previous knowledge about the data (Rendón et al. 2011) while the internal criteria is based on the information from the clustering solution. In this thesis, only the internal CVAIs are considered.

The internal CVAIs are based on measures of within-cluster (intra) and between-cluster (inter) separability, taking into account either one or both of these measures. Proper CVAIs measure how well the general goal of clustering - high similarity within clusters, i.e., small intra, and low similarity between clusters, i.e., large inter - is reached, when the iterative relocation algorithms (see Chapter 3.4) only greedily decreases the clustering error locally. Generally indices measuring both intra and inter are taken as their division – if intra is as numerator, the best index value is at the minimum, while for inter as numerator it is at the maximum.

There exists a lot of different internal CVAIs and even several forms of them. For example, a well known CVAI, Davies Bouldin, was first introduced in (Davies and Bouldin 1979), a new version, Davies Bouldin*, was introduced in (Kim and Ramakrishna 2005), and a slightly different version is also implemented in MATLAB (Documentation 2015). In Chapter 5, we present our results for comparing 43 different CVAIs with 12 synthetic data sets, and in this chapter we will present the ones that were found to work best.

Some of the simplest CVAIs are based only on the intra measure. These include, for example, the Ball Hall (BH) (Ball and Hall 1965) and so-called knee-point/elbow methods (Thorndike 1953). The elbow methods are based on choosing the number of clusters based on the point where the within-cluster error bends. The elbow point can be determined either by plotting the values and visually observing the plot or by finding the maximum difference between two points.

Ball Hall defines intra as the mean of clustering error, J_k, divided by the size of the cluster:

$$BH = Intra = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}J_k^*, \tag{3.6}$$

where n_k is the size of cluster k and the optimal solution is to minimize this within-cluster separation.

kCE (Jauhiainen and Kärkkäinen 2017) is an index based on the clustering error J and the number of clusters, so it only considers the intra measure. The index is defined as

$$kCE = K \times J^*. \tag{3.7}$$

This measures if adding the number of clusters by one pays off or not. So having two prototypes should reduce the clustering error by two, having three should decrease it by three etc. The optimal number of clusters is found at the minimum index value.

The Ray Turi (RT) index (Ray and Turi 1999) is based on both intra and inter measures. The intra is defined as mean of the clustering error J, inter as the minimum comparative distances (with respect to the clustering error, see (3.5)) between prototype centers and the index as

$$RT = \frac{Intra}{Inter}, \tag{3.8}$$

where

$$\begin{cases} Intra = \dfrac{1}{N} \times J^* \\ Inter = min(\|\mathbf{c}_k - \mathbf{c}_{k'}\|_p^q) \end{cases} \tag{3.9}$$

and $k, k' \in \{1, ..., K\}$, $k \neq k'$. The optimal solution is achieved by minimizing this index – the smaller the within-group and the larger the between-group separability, the better the clustering results.

Another popular index, Calinski Harabasz (CH) (Caliński and Harabasz 1974), defines the intra with the clustering error J and the inter with the sum of l_p^q-distances between the prototype centers and the whole data center, weighted with the size of the cluster. The index is defined as

$$CH = \frac{Inter}{Intra}, \tag{3.10}$$

where

$$\begin{cases} Intra = (K-1) \times J^* \\ Inter = (N-K) \times \sum_{k=1}^{K} n_k \|\mathbf{c}_k - \mathbf{c}_0\|_p^q \end{cases} \tag{3.11}$$

and n_k is the size of cluster k and $\mathbf{c}_0$ is the center of whole data. The optimal solution here is, on the contrary to Ray Turi, the maximum value. Notice the close relation of this intra measure with the kCE index given in (3.7).

Davies Bouldin (DB) is based on a ratio of the intra and inter measures. The inter is defined as the distance between two cluster centers, $\mathbf{c}_k$ and $\mathbf{c}_{k'}$, and intra as sum of the average distances between each point in clusters k and k' to its cluster center. The index considers the worst ratio between the measures, and the actual value is taken as average sum over these ratios:

$$DB = \frac{1}{K} \sum_{k=1}^{K} \max_{k \neq k'} \frac{Intra(k,k')}{Inter(k,k')}, \tag{3.12}$$

where

$$\begin{cases} Intra = \frac{1}{n_k} J_k^* + \frac{1}{n_{k'}} J_k^* \\ Inter = \|\mathbf{c}_k - \mathbf{c}_{k'}\|_p^q. \end{cases} \tag{3.13}$$

The optimal solution with Davies-Bouldin index is found in the minimum value with respect to K.

The PBM index (acronym from Pakhira, Bandyopadhyay, and Maulik) (Pakhira, Bandyopadhyay, and Maulik 2004) defines the intra measure with the clustering error J and the inter with the l_p^q-distances between cluster centers and distances between data points and the center of the whole data. The index is defined as following

$$PBM = \left(\frac{Inter}{Intra}\right)^2, \tag{3.14}$$

where

$$\begin{cases} Intra = K \times J^* \\ Inter = max(\|\mathbf{c}_k - \mathbf{c}_{k'}\|_p^q) \times D \end{cases} \tag{3.15}$$

and $D = \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{c}_0\|_p^q$, so the sum of distances of all points to the l_p^q-center of the whole data. The maximum value of the index indicated the best number of clusters. Again notice that intra of PBM is exactly the definition kCE in (3.7).

Wemmert-Gançarski is another index based on the ratio of intra and inter measures. It defines the intra measure as the distance between the point $\mathbf{x}_i$ and the center of the cluster it belongs to. The inter is defined as the minimum distance of the point to the centers of all the other clusters. The ratio between these measures is considered and the actual index is the weighted mean of the mean ratios in each cluster. If the mean ratio in a cluster is greater than 1, it is ignored, otherwise its complement to 1 is considered:

$$WG = \frac{1}{N} \sum_{k=1}^{K} max\Big(0,\ n_k - \sum_{i \in I_k} \frac{Intra(i)}{Inter(i)}\Big), \tag{3.16}$$

where I_k is the set of point that belong to cluster k and the intra and inter values are defined as

$$\begin{cases} Intra(i) = \|\mathbf{x}_i - \mathbf{c}_k\|_p^q \\ Inter(i) = \min\limits_{k \neq k'} \|\mathbf{x}_i - \mathbf{c}_{k'}\|_p^q. \end{cases} \tag{3.17}$$

Silhouette (Rousseeuw 1987) index is based on silhouette values, measuring the similarity of a point, $\mathbf{x}_i$, to points in the same cluster, when compared to points in other clusters. The intra measure is the l_p^q-distance between points in cluster and inter is the minimum average l_p^q-distance from a point in cluster to points in a different cluster:

$$\begin{cases} Intra(i) = \dfrac{1}{n_k - 1} \sum_{j,i \in I_k, i \neq j} \|\mathbf{x}_i - \mathbf{x}_j\|_p^q \\ Inter(i) = \min\limits_{k \neq k'} \Big(\dfrac{1}{n_k} \sum_{i \in I_k, j \in I_{k'}} \|\mathbf{x}_i - \mathbf{x}_j\|_p^q \Big) \end{cases} \tag{3.18}$$

Then, for each point $\mathbf{x}_i$, a value $s(i)$, indicating the silhouette width of the point, is formed as follows:

$$s(i) = \frac{Intra(i) - Inter(i)}{max(Inter(i),\, Intra(i))}. \tag{3.19}$$

The silhouette value, $s(i)$, ranges from -1 to 1. A high value indicates that $\mathbf{x}_i$ is well-matched to its own cluster, but poorly to the neighboring clusters. If most points have a high value, the clustering partition is appropriate. So finally, the actual index value is taken as the mean of the mean silhouettes through all clusters:

$$SILH = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i \in I_k} s(i). \tag{3.20}$$

The maximum index value proposes the number of clusters. Compared to other CVAIs, silhouette is a lot slower to calculate. It goes through the data twice and so the performance is $O(N^2)$. Generalized Dunn indices (GDI) (Bezdek and Pal 1998), generalizations from the original Dunn index (Dunn 1974), are another example of CVAIs that go through the data more than once, and therefore are more complex than most of the CVAIs.

3.7 Summary

Clustering is unsupervised classification, where the possible classes are not known in prior, but are determined from the data. The purpose is to form the clusters such that the members in a cluster as as similar to each other as possible and as dissimilar to the members in other clusters as possible (Jain, Murty, and Flynn 1999). The similarity can be measured by, for example, distance or density. The clustering task can be very ambiguous due to different similarity measures, noise in data and the decision whether the clustering is considered locally or globally. Therefore, there is no clear definition for a cluster and the ”clusters are, in

large part, on the eye of the beholder" (Estivill-Castro 2002). The amount of clusters present in the data can be evaluated with CVAIs and some best known and well working ones were introduces in this chapter. Note that the introduction here is novel, because the original definitions of the cluster indices have been given only in relation to the squared Euclidean distance, i.e., in the context of k-means-type of algorithms for $p = q = 2$.

4 Measuring and quantifying physical activity

Caspersen et al. define physical activity as "any bodily movement produced by skeletal muscles that requires energy expenditure" (Caspersen, Powell, and Christenson 1985). It is related to the whole span of individuals life and it commonly changes its form and decreases with age (Hallal et al. 2012). It can happen at school, at work, during transportation or during free time. The activity at free time can be further divided into housework, sports, and other activities (Caspersen, Powell, and Christenson 1985). Another division is to separate activity into voluntary and compulsory.

For data generation about the form, intensity and amount of physical activity, many sensors and other techniques have been developed and utilized. Different activity monitors have nowadays become very common and capable of measuring different components of physical activity, as well as sleep and recovery from training (Evenson, Goto, and Furberg 2015). As a growing number of people own even multiple monitors and sensors, an emerging trend, quantified self, has been born (Swan 2013). According to Swan, quantified self denotes any individual measuring biological, physical, behavioral, or enviromental information of themselves by self-tracking. As this measured information creates vast streams of data and many monitor manufactures and activity applications enable loading these to the internet, the consept of Internet of Humans (IoH) has been recognised as a part of the Internet of Things (IoT) (Arbia et al. 2015).

The way of activity measuring has to be selected case dependently – according to the physical activity component assessed, characteristics of the target population, and feasibility of the sensor in terms of, e.g., cost (Butte, Ekelund, and Westerterp 2012). Next we introduce some components of physical activity that are often considered and following some well known and widely used sensors for measuring.

Energy expenditure:

Energy expenditure is measured in kilocalories (Caspersen, Powell, and Christenson 1985). The total daily energy expenditure (TDEE) consists of basal energy expenditure (BEE), diet-unduced thermogenesis (DIT) and activity-related energy expenditure (AEE) (Bonomi

2010). TDEE can be measured with indirect calorimetry, such as the doubly labeled water method (Schoeller and Van Santen 1982), that is considered as a gold standard technique (Bonomi 2010).

Activity intensities:

The intensity of activity is often divided into categories of sedentary, light, moderate and vigorous. These are commonly defined based on the energy expenditure expressed as METs – light < 3 METs, moderate 3-6 METs, vigorous > 6 METs (Ainsworth et al. 2000). The division into these categories is also often made by the counts per minute (CPM). Freedson et. al. classified the activity into four intensity levels by counts per minute. Upper limits were 1951 for light, 5724 for moderate, 9498 for hard and counts that were greater than 9498 corresponded to very hard intensity level (Freedson, Melanson, and Sirard 1998). Also, the time spent in moderate to vigorous physical activity (MVPA) is a measure used lot in the research of activity.

Similar categories have been constituted separately for children, for example using Acti-Graph by Evenson et al. (Evenson et al. 2008). In their study the upper limits were 100 for sedentary, 2292 for light, 4008 for moderate and counts that were greater corresponded to vigorous intensity activity.

4.1 Accelerometers

Accelerometers are devices that measure body movements in terms of acceleration (Chen and Bassett 2005). Most of the accelerometers are small Micro-Electro-Mechanical Sensors (MEMS). There is a small proof mass in the MEMS accelerometers and it measures acceleration from the displacement of the mass (Yang and Hsu 2010). Acceleration can be measured in one, two, or three orthogonal axes and is quantified often in terms of standard gravity (g). The bandwidth of the accelerometer is measured in Hertz ($1/s$) and it denotes the frequency of how often readings are stored in a second.

With human activity recognition, meaning the classification of different activities, triaxial accelerometers are one of the most broadly used sensors (Lara and Labrador 2013). Infor-

mation about the intensity of activity can also be gained from acceleration data, which is a factor in accelerometers becoming so widely used in the field of physical activity.

Accelerometers can be build upon many different technologies, capacitive, piezoelectric and piezoresistive ones being the most popular.

Piezoelectric

Piezoelectric accelerometers are relatively small, lightweight, and used in most accelerometry-based physical activity monitors (Chen and Bassett 2005). These accelelerometers can not detect the orientation of body parts or static activities because they are equipped with piezoelectric sensors that measure acceleration due to movement and cannot measure static forces (Bonomi 2010).

There are two main types of piezoelectric accelerometer; one with a cantilever beam and an other with integrated circuit (IC) chip. The first kind has a beam, piezoelectric element, and a seismic mass, and the second kind has an integrated chip sensor instead of the beam. When seismic mass detects acceleration it causes bending to the beam or tension and compression to the integrated chip sensor which in turn is sensed by the piezoelectric element as deformation (Chen and Bassett 2005).

Capacitive

Capacitive sensors have become popular because they offer high sensitivity, good long term stability, and are low power (Bao 2000). They output a voltage dependent on the distance between planes. When one of these planes is moved it changes the electical capacity of the system. With capacitive accelerometers it is possible to detect static postures like lying down, sitting, and standing, that can be important components in studies, which for one has made them more popular in human activity research (Bonomi 2010).

Piezoresistive

Piezoresistive accelerometers measure the electrical resistance of a material when mechanical stress is applied to it. Similarly to piezoelectric sensors it has a beam and when acceleration is directed to it, the inertial force of the mass bends the beam (Bao 2000). This bending

causes change in the resistance of the piezoresistor. Piezoresistive sensors are sensitive and also have the ability to detect static postures like capacitive accelerometers (Bonomi 2010).

4.1.1 ActiGraph and counts

ActiGraph (Pensacola, FL) is an activity monitor based on a three axis accelerometer, that can be worn on the wrist, waist, ankle, or thigh. They have been widely used in the research of physical activity (Sasaki, John, and Freedson 2011) and on their homepage it is said that "ActiGraph accelerometry monitors are among the most widely used and extensively validated devices of their kind" [1].

The initial voltage signals from most accelerometry sensors are, after filtering and amplifing, sampled at a fixed frequency and converted from analog signal to digital numbers, which are called "raw counts" (Chen and Bassett 2005). These are however not the same counts that most of the current physical activity monitors output. These "raw counts" are further processed with different approaches, for example, one of the most common is to integrate over a time interval, usually one minute (Bonomi 2010).

ActiGraph uses counts that are defined on their support page[2] as "a result of summing post-filtered accelerometer values (raw data at 30Hz) into epoch "chunks." The value of the counts will vary based on the frequency and intensity of the raw acceleration."

4.2 Gyroscopes

Gyroscopes are rotation sensors that measure the angular velocity, position, and orientation of a moving object. Similarly to accelerometers, most used gyroscopes are Micro-Electro-Mechanical Sensors (MEMS) because they are small, inexpensive, reliable, and low power (Avci et al. 2010). They can be integrated together with accelerometers to estimate the orientation of a device better and can hence increase the accuracy of human activity recognition (Ustev, Durmaz Incel, and Ersoy 2013).

1. http://actigraphcorp.com/
2. https://actigraph.desk.com/customer/en/portal/articles/2515580-what-are-counts-

4.3 Pedometers

Pedometers are small, inexpensive devices that can count steps and estimate distances as well as energy expenditure. The effect of steps can be detected with a switch mechanism like a spring-loaded mass inside the device (Yang and Hsu 2010). Pedometers are quite accurate at step counting but other measurements can vary a lot between different pedometers. One major disadvantage is that the intensity of the activity can't be quantified using pedometers. Some pedometers allow one to enter the stride length and those work reasonably accurately with step counting at normal walking speeds but overestimate the steps at slow speeds and underestimate them at high speeds (Butte, Ekelund, and Westerterp 2012).

4.4 Heart rate monitors

Heart rate sensors are widely used for measuring the human activity. Heart rate is relatively easy to measure and the monitors are of reasonable cost. Compared to accelerometers, they can give more precise information about the intensity of activities where the speed does not reflect the intensity (Achten and Jeukendrup 2003). A good example is exercising with a stationary bicycle when the accelerometer might detect almost no activity at all, especially if placed on the wrist, so the heart rate sensor becomes very useful in getting more appropriate information about the intensity.

Heart rate sensors are often also used together with accelerometers. Because an accelerometer measures only the movement it detects, heart rate sensors can give very useful additional information. Especially the estimation of energy expenditure and the intensity of the physical activity are more valid with the combined usage of accelerometers and heart rate sensors. Intille et al., (2012) recommend that the engineers and device developers should combine heart rate monitoring with other sensing technologies more in the future.

4.5 Vision sensors

Vision sensors have been used commonly by researchers in the past (Pentland 2000). They are especially suitable for security and interactive applications (Lara and Labrador 2013),

as well as for human activity recognition and detecting and identifying people. In sports, cameras can be exploited for team sports to track and measure how much and in what ways the players are moving in the field or rink, and that information can then be further used for optimizing and creating new tactics (D'Orazio and Leo 2010).

Using vision sensing has its problems – privacy and maintaining issues, especially if they are being used in individuals free-living conditions. Cameras need a lot of technical support and the attaching of the devices is not unambiguous. As other sensor concurrently progress and increase, the visions sensor are becoming less used in some fields of human activity, such as activity recognition. On the other hand, they are also becoming more general in other field, such as physical activity games.

4.6 Global Positioning System

Global Positioning System (GPS) can provide information on a person's location, enviroment, mode of transportation, and speed with a satellite-based system (Butte, Ekelund, and Westerterp 2012). The information about the location can be helpful with activity recognition (Lara and Labrador 2013). GPS has been integrated in many sports applications and activity monitors, but can technically be used only in outside activities since the satellites do not work properly without a clear access. Some other downsides are that the GPS can be computationally expensive and can violate ones privacy.

4.7 Physical activity data applications

Multiple applications utilizing the above sensors have been developed. These applications are from a variety of domains, for example, medical or sports. The data derived from the sensors can be processed following the KDD process - preprocessing is almost always necessary with any kind of data, while transformation can refer to summing up the accelerometer counts over time, dimension reduction, or scaling of the data. With data mining the actual information, that the application needs, is derived.

4.7.1 Human activity recognition

In activity recognition the goal is to recognize common human activities in real life settings (Kim, Helal, and Cook 2010). The process of recognizing can be summarised as "determining a target set of activities, collecting sensor readings, and assigning sensor readings to the appropriete activites" (Incel, Kose, and Ersoy 2013). These sensor readings can be from, for instance, vision sensors, accelerometers, or gyroscopes. The assignment of a reading is most often done by classification and due to the supervised nature of the process, training dataset with acitivity labels is required (Preece et al. 2009).

The steps of the activity recognition process can be divided into data collecting, segmentation, feature extraction, and classification (Preece et al. 2009; Duda, Hart, and Stork 2012). Data collecting is the stage where the data about physical activity is being collected by a monitor and then labeled with the corresponding category labels. Also, many datasets for human activity recognition are available, such as " Activity Recognition system based on Multisensor data fusion (AReM) Data Set"[3].After the data is collected and labeled, usually some preprocessing is done, e.g., for noise removal (Incel, Kose, and Ersoy 2013). In classification, corresponding to the data mining step of the KDD process, the activity is classified into different predefined categories, for example running, walking, sitting, gardening or driving a car, using classification methods (see Chapter 2.4.3). Segmentation and feature extraction correspond the transformation step in the KDD process and are introduced next.

Segmentation

Segmentation is needed because finding meaningful and useful information from a continuous stream of data can be difficult (Avci et al. 2010). The purpose is to identify those segments of the preprocessed data streams that are likely to contain information about activities (Bulling, Blanke, and Schiele 2014).

Windowing techniques are the most used segmenation methods with activity classification Preece et al. 2009) and, of those, sliding windows are the most widely used, because they are simple, intuitive, and easy to implement (Avci et al. 2010). Data inside a window with

3. http://archive.ics.uci.edu/ml/datasets/Activity+Recognition+system+based+on+Multisensor+data+fusion+%28AReM%29

certain length is a segment and this window is being moved over the data to get the segments for further processing. The step size that the window is being moved has to be selected case dependently as well as the window length. Some comparisons with different window lengths with human activity recognition have been made in (Mannini et al. 2013; Bulling, Blanke, and Schiele 2014;Huynh and Schiele 2005).

Feature extraction

The purpose of feature extraction is to find the informative characteristics of a data segment that accurately represent the original data (Incel, Kose, and Ersoy 2013). The signals are being reduced into features that distinguish the activities as well as possible (Bulling, Blanke, and Schiele 2014) and, in an ideal situation, the feature extraction works so that it makes the job of a classifier trivial (Duda, Hart, and Stork 2012).

The most simple features are signal statistics, such as mean, variance, or root mean square, that can be extracted quite easily and automatically from the data. Other widely used features are the frequency-domain features that focus on the periodic properties in the data (Avci et al. 2010). These can be derived using, for instance, Fast Fourier transformations or wavelet analysis.

4.7.2 Gait analysis

Walking is one of the most common forms of physical activity and it has a necessary role in our everyday lifes (Rueterbories et al. 2010). The term gait is used to describe the way of walking and gait analysis is the examination of the pattern of walking (Whittle 2014). With gait analysis, a lot of useful health related information can be gained.

Gait analysis has been approached in many different ways. The most ordinary way is to go to a clinic, where a healthcare professional, for example physiotherapist, visually observes the gait. Laboratory testing in turn is usual among top athletes, but not widespread in the research of locomotor disorders (Simon 2004). However, many wireless systems have been developed to be used outside laboratories, using, for example, vision sensors (Stone and Skubic 2011), gyroscopes (Tong and Granat 1999), accelerometers (Hartmann et al. 2009), and the combination of multiple sensors (Bamberg et al. 2008).

4.7.3 Monitoring and medical diagnosis

Increasing health-care costs of the aging population have become a concern in many countries and the potential of human activity recognition in eldery care has been a subject of interest (Jiang et al. 2008; Tapia, Intille, and Larson 2004; Najafi et al. 2003). Monitoring of elderly and their activity with different sensors from outside their free-living conditions makes it possible for them to stay at their homes more safely. With monitoring and activity recognition techniques, any abnormal behaviour could be detected and, furthermore, based on recognized activities, reminders for necessary activities could be included if needed (Avci et al. 2010). Moreover falling, a great risk for the old people, can be detected by someone from without and if no movement follows the falling, help can be alerted. In addition to human activity recognition, also gait analysis can be used when trying to detect falling. For example, foot clearance is an important gait parameter when considering the risk of falling and algorithms for estimating it have been developed (Morales Gonzalez 2015). These similar monitoring principles could be used with disabled people or even with children.

Human activity monitoring and gait analysis can also be used for diagnosing diseases such as dementia or Parkinson's disease. In addition, some information about the phase and pace of the conditions can be gained. In Parkinson's disease the most common sympton is tremor, occuring in almost every patient (Jankovic 2008). The disase and its pace can be recognised by it and a system for diagnosing and predicting the pace of the Parkinson's disease has been developed (Atif and Serdaroglu 2012). Gait abnormalities in turn have been linked with potential dementia (Marquis et al. 2002; Davis 1988) and on that account a system to analyse the gait process has been presented by Lofti et al. (Lotfi, Nguyen, and Langensiepen 2015).

4.7.4 Activity and sports

Another popular field for utilizing the physical activity monitors is sports. This can be considering competitive sports or just daily sport activities by non-athletes. Many commercial monitoring systems have been developed specifically for sports and even for certain kinds, such as running or cycling. They give useful information during the exercise (heart rate,

stride ratecadence when running, or speed and distance), as well as over a longer time span by tracking e.g. ones sleep, training, and general activity.

In competitive sports, optimizing the performance is very important and for example running performance and foot-ground contact have been analysed in a study by Gasser (Gasser 2014). In addition the heart rate variability has been under an increasing interest by researchers. Overtraining in athletes can be prevented and detected (Achten and Jeukendrup 2003) and adaption to endurance training can be assessed (Vesterinen et al. 2013) with heart rate variability.

4.8 Activity monitors

The MEMS accelerometers and gyroscopes can be used e.g in mobile phones, computers, tablets, or in devices designed specifically for physical activity measuring. These small devices often include the accelerometer, battery, some memory, and a channel for transferring the data and sometimes the gyroscope. They write down the accelerometer readings and those can be further used in researches.

Modern mobile phones have many sensors in them including accelerometers and gyroscopes. Also the GPS can be utilized for measuring the speed and distance traveled with mobiles. A variety of free and commercial applications have been developed for getting information about activity with mobile phones. The advantage with using mobiles is that most people already carry those with them almost all the time. Many studies have been made about activity recognition from the mobile phone accelerometer data. Baya et al. (2014) studied six physical activity patterns: slow walking, fast walking, running, stairs-up, stairs-down, and aerobic dancing using an Android smartphone (Bayat, Pomplun, and Tran 2014), reaching an overall accuracy rate of 91.15% for recognition. Bremez et. al. used Nokia N95 for classifying walking, stairs-up, stairs-down, standing-up, sitting-down, and falling and got results that were quite accurate, between 70 to 90% for different activities (Brezmes, Gorricho, and Cotrina 2009).

4.8.1 Consumer-based physical activity monitors

Multiple commercial activity monitors have been developed, becoming very affordable and common for the masses. Some well known manufacturers are Fitbit, Nike, BodyMedia, Polar, Garmin, Jawbone, and ActiGraph, having multiple models for different uses. While some models are developed more for general activity monitoring, some are specialised in monitoring certain sports. Polar for example has monitors developed specifically for running and cycling, Garmin has ones developed for swimming and golf. While most of commercial monitors are worn on the wrist, some are meant to be worn on the hip, for example, ActiGraph and Fitbit One, and some can be worn on the ankle or around the neck.

4.8.2 Validity of consumer-based physical activity monitors

The validity of physical activity monitors can vary and the subject has been studied a lot. It has been found that many accelerometer based monitors tend to underestimate the energy expendidure both during exercise and in free-living conditions (Crouter, Churilla, and Bassett Jr 2006; Hendelman et al. 2000; Carmines and Zeller 1979).

The validity of nine different monitors (Fitbit One, Fitbit Zip, Jawbone UP, Misfit Shine, Nike Fuelband, Striiv Smart Pedometer, Withings Pulse, BodyMedia SenseWear, and ActiGraph GT3X+) was examined by Fergunson et. al. (Ferguson et al. 2015). They had 21 healthy adults wearing the monitors for 48 hours in their free-living conditions. The monitors showed strong validity for measuring steps and sleep duration but only moderate validity for moderate to vigorous physical activity (MVPA) time. All devises in the study underestimated the total daily energy expenditure (TDEE), compared to the reference device (SenseWear).

Nelson et al., (2016) tested the accuracy of the Fitbit One, Zip, and Flex and Jawbone UP24 for estimating the energy expenditure and steps for specific activities. All monitors in their study severely underestimated energy expenditure during cycling. They also concluded that the measurement for steps was accurate but that the monitors should be used cautiously for estimating the energy expenditure.

In their study Lee et al. (Lee, Kim, and Welk 2014) examined the validity of energy expenditure from variety of monitors, under free-living conditions. BodyMedia FIT armband

was worn on the left arm, DirectLife monitor around the neck, Fitbit One, Fitbit Zip, and ActiGraph worn on the belt, and Jawbone Up and Basis B1 Band monitor on the wrist. Sixty healthy adults wore the monitors and completed a routine including 13 different activities. For overall group comparisons, the mean absolute percent error values were 9.3%, 10.1%, 10.4%, 12.2%, 12.6%, 12.8%, 13.0%, and 23.5% for the BodyMedia FIT, Fitbit Zip, Fitbit One, Jawbone Up, ActiGraph, DirectLife, NikeFuel Band, and Basis B1 Band, respectively. The results clearly favored the BodyMedia FIT armband, but promising results were also observed with the Fitbit Zip.

Nowadays most commercial activity monitors do combine the accelerometer and heart rate monitor to improve the accuracy in measuring both physical activity and the energy expenditure.

4.9 Summary

Physical activity and its different components can be measured with a variety of different sensors, that have become very accessible and widely used. People measuring and sensoring themselves, their sleep, heart rate etc., with different activity monitors has lead to an emerging trend, quantified self. As this data is more often uploaded to the internet, the concept of Internet of Humans has arisen (Arbia et al. 2015). Multiple sensors for measuring the physical activity were introduced in this chapter, including accelerometers, heart rate sensors, pedometers, and gyroscopes. As more and cheaper customer-based physical activity monitors are coming to the markets, their validity was also discussed. In general, many of these monitors tend to underestimate the energy expenditure, but the validity also varies a lot between different monitors. Therefore the monitor should always be chosen case-dependently - in recreational sports the validity does not need to be as good as with, for example, medical use.

5 CVAI tests

This chapter recapitulates the work already presented in (Jauhiainen and Kärkkäinen 2017), where a set of CVAIs and their different available implementations for R and MATLAB platforms were compared to choose the most suitable ones for the clustering of students activity and sedentary behaviour in Chapter 6. The performance of the indeces was tested with noisy datas as well as datas of different dimensions, densities, and distributions. Altogether 43 CVAIs were tested with 12 synthetic datasets. The best performing CVAIs out of these were introduced in Chapter 3.6.

5.1 Test datas

The comparisons were run with 12 synthetic datasets, of which four were self simulated. These included four 2D S-datasets (see Fig. 10), described in (Fränti and Virmajoki 2006), and four higher dimensional Dim-datasets from http://cs.uef.fi/sipu/datasets/. The S-sets have 15 known centers with increasing noise so that S_4 data has the most noise. The Dim-datasets are of dimension 32, 64, 128, and 256 and have 16 known centers.

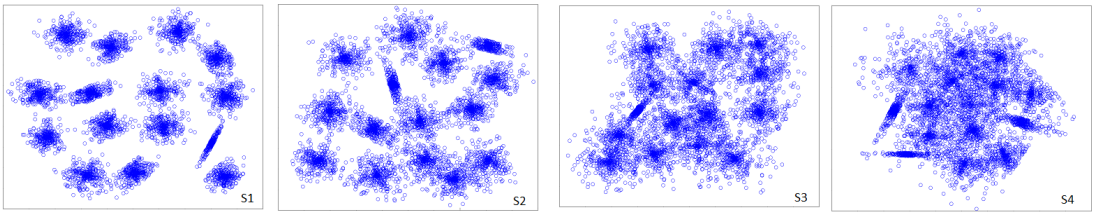

Figure 10. The S-datasets with 15 clusters and increasing noise.

Additional simulated datasets were also created to test the CVAIs with some more specific cases. For example, S1D2 is a dataset having clearly just one cluster, while S2D2 (see Fig 11) has two clusters close to each other, with some additional noise. The third simulated data, S5D2 (see Fig. 11), was implemented so that it has 10% additional noise with five clusters of which two are harder to detect being smaller, more sparse, and situated close to a bigger cluster. S5D10 is similar to this but in 10D.

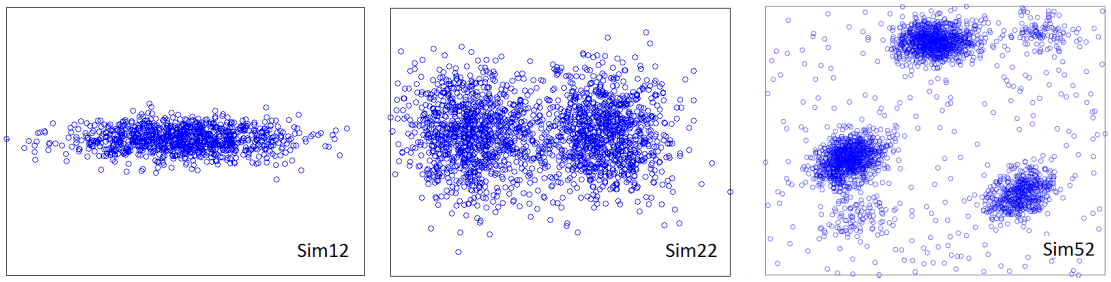

Figure 11. From left to right: S1D2, S2D2, and S5D2.

5.2 Methods

First of all, we tried to identify the most actively used or recent CVAI packages from R and Matlab platforms. To this end, CVAI implementations from five different packages were used, including threeR packages, `NbClust`[1] (`P1`), `cclust`[2] (`P2`), and `clusterCrit`[3] (`P3`) with 30, 15, and 27 implemented CVAIs, respectively. In MATLAB's function evalclusters (`P4`) we applied three CVAIs (Davies-Bouldin, Calinski-Harabasz, and Silhouette), while 10 CVAIs from the `Cluster Validity Analysis Platform`, `CVAP` (`P5`), downloaded from MATLAB's file exchange center, were tested.

The datas were first min-max scaled into the range of $[-1,1]$ and then clustered using the k-means algorithm. This clustering was done beforehands in MATLAB with all packages but the `NbClust`, because with it the clustering had to be done in R at the same time as the CVAIs were calculated. The k-means algorithm in MATLAB was repeated 1000 times and the solution with smallest clustering error was selected as the final result. The CVAIs were computated for K =2–20.

5.3 Results

The suggested number of clusters highly varied between different CVAIs and even between their implementations. The results for 17 best CVAIs can be seen in Table 2. Hyphen in the table means that there was no implementation of that CVAI in the package or that the

1. https://cran.r-project.org/web/packages/NbClust/NbClust.pdf
2. https://cran.r-project.org/web/packages/cclust/cclust.pdf
3. https://cran.r-project.org/web/packages/clusterCrit/clusterCrit.pdf

calculation failed (producing NaN, inf etc.). Row "Correct" measured the difficulty of a data by counting the number of correctly determined number of clusters with at least one implementation. The number of correct propositions from an implementation of an index is given in column "Correct", where next to the last shell provides the median of correct propositions over all packages.

We noticed that the clustering in MATLAB improved the performance of the CVAIs in the two R-packages `P2` and `P3` compared to if the clustering was also done in R. That might be one of the reasons why the `P1` was by far the worst package – with it the clustering had to be done with k-means built in R. Also the clustering with this package was done with fewer repetitions, the default in R being 10.

Overall the CVAIs of MATLAB worked a lot better, but few ones in R's `P3` also performed well. These include the Pakhira-Bandyopadhyay-Maulik (PBM), Wemmert-Gançarski, and Calinski Harabasz. In addition to these three indeces in `P3`, the best CVAIs include Davies Bouldin in `P4` as well as Silhouette and Calinski Harabasz in `P3` and `P4`. All in all `P4` was the best performing package, with the median on correct propositions being nine out of 12 datasets. This is probably due to carefully selected CVAIs in the first place, since it is a commercial package. Calinski Harabasz performed well with all packages but the worst, `P1`. The only index that suggested the right number of clusters in all the cases, was kCE.

5.4 Conclusion

Usually when new CVAIs are introduced, the paper also includes an experimental evalution of multiple indices, typically concluding the proposed index as the best one. In addition, some methods that execute both the clustering and the analysis of the number of clusters have been developed, such as the Viral Clustering (VA) algorithm in (Petrosyan and Proutiere 2016), where it was compared to seven CVAIs and concluded as the best performing one.

Moreover, eight CVAIs were compared in (Liu et al. 2010). Most suggested the correct number of clusters with 5% additional noise, different densities, and skewed distributions, while only three of them were able to recognize closed subclusters. Sdbw was the only CVAI that suggested the right number of clusters for all data sets. Often no single CVAI dominates

Table 2. Numbers of clusters suggested by the CVAIs

P1,P2,P3 P4,P5	*S1*	*S2*	*S3*	*S4*	*D32*	*D64*	*D128*	*D256*	*S5D2*	*S5D10*	*S2D2*	*S1D2*	Correct
Davies Bouldin	**15**,**15**,17 **15**,**15**	14,**15**,16 **15**,14	10,**15**,**15** **15**,**15**	14,17,14 14,**15**	15,**16**,**16** **16**,17	20,**16**,**16** **16**,19	18,**16**,**16** **16**,18	18,**16**,**16** **16**,18	3,3,4 3,3	3,3,3 3,2	**2**,16,7 **2**,14	18,16,18 20,15	2,7,5 **8**,3
Calinski Harabasz	16,**15**,**15** **15**,**15**	14,**15**,**15** **15**,**15**	19,**15**,**15** **15**,**15**	**15**,**15**,**15** **15**,**15**	17,**16**,**16** **16**,**16**	19,**16**,**16** **16**,**16**	18,**16**,**16** **16**,**16**	17,**16**,**16** **16**,**16**	3,3,3 3,3	3,3,3 3,4	**2**,**2**,**2** **2**,**2**	6,6,6 6,6	2,**9**,**9** **9**,**9**
Silhouette	**15**,-,**15** **15**,**15**	14,-,**15** **15**,**15**	**15**,-,**15** **15**,**15**	**15**,-,14 **15**,**15**	15,-,- **16**,**16**	19,-,- **16**,**16**	18,-,- **16**,**16**	17,-,- **16**,**16**	3,-,3 3,3	3,-,3 3,4	**2**,-,**2** **2**,**2**	17,-,18 18,18	4,-,4 **9**,**9**
Hartigan	7,20,**15** -,2	14,20,4 -,2	4,20,4 -,2	3,20,3 -,2	15,**16**,**16** -,2	18,**16**,**16** -,2	18,**16**,**16** -,2	17,**16**,**16** -,2	3,20,3 -,2	3,2,3 -,3	4,16,4 -,**2**	3,16,3 -,2	0,4,**5** -,1
Dunn	9,-,**15** -,**15**	9,-,**15** -,**15**	5,-,20 -,4	10,-,**15** -,4	3,-,**16** -,**16**	5,-,**16** -,**16**	5,-,**16** -,**16**	8,-,**16** -,**16**	3,-,3 -,3	19,-,19 -,4	18,-,20 -,**2**	19,-,13 -,3	0,-,**7** -,**7**
Cidx	17,-,**15** -,20	**15**,-,**15** -,20	20,-,20 -,20	16,-,20 -,20	15,-,**16** -,-	19,-,**16** -,20	18,-,14 -,20	17,-15 -,-	17,-,3 -,20	4,-,7 -,2	3,-,20 -,20	17,-,20 -,20	1,-,**4** -,0
Rubin	**15**,**15**,**15**	14,**15**,**15**	19,**15**,**15**	14,**15**,**15**	15,3,**16**	19,3,13	18,3,19	17,3,19	3,19,19	3,17,17	4,12,12	17,15,15	1,4,**5**
Ray Turi	-,-,**15**	-,-,**15**	-,-,4	-,-,13	-,-,**16**	-,-,**16**	-,-,**16**	-,-,**16**	-,-,3	-,-,3	-,-,**2**	-,-,6	-,-,**7**
Sdbw	**15**,-,**15**	19,-,**15**	20,-,2	19,-,2	20,-,-	20,-,-	20,-,-	20,-,-	**5**,-,3	20,-,-	20,-,**2**	20,-,2	2,-,**3**
GenDunn	-,-,**15**	-,-,**15**	-,-,20	-,-,**15**	-,-,**16**	-,-,**16**	-,-,**16**	-,-,**16**	-,-,2	-,-,19	-,-,20	-,-,13	-,-,**7**
Gamma	-,-,**15**	-,-,**15**	-,-,20	-,-,20	-,-,**16**	-,-,**16**	-,-,14	-,-,15	-,-,**5**	-,-,7	-,-,20	-,-,20	-,-,**5**
G+	-,-,**15**	-,-,**15**	-,-,20	-,-,20	-,-,**16**	-,-,**16**	-,-,14	-,-,15	-,-,**5**	-,-,7	-,-,20	-,-,20	-,-,**5**
PBM	-,-,**15**	-,-,**15**	-,-,5	-,-,4	-,-,**16**	-,-,**16**	-,-,**16**	-,-,**16**	-,-,**5**	-,-,**5**	-,-,**2**	-,-,4	-,-,**9**
WemGan	-,-,**15**	-,-,**15**	-,-,**15**	-,-,**15**	-,-,**16**	-,-,**16**	-,-,**16**	-,-,**16**	-,-,3	-,-,3	-,-,**2**	-,-,20	-,-,**9**
Xie Beni	-,-,**15**	-,-,**15**	-,-,20	-,-,16	-,-,**16**	-,-,**16**	-,-,**16**	-,-,**16**	-,-,3	-,-,11	-,-,20	-,-,3	-,-,**6**
CXu	-,**15**,-	-,**15**,-	-,16,-	-,16,-	-,2,-	-,2,-	-,2,-	-,2,-	-,**5**,-	-,19,-	-,**2**,-	-,3,-	-,-,**4**
Ssi	-,13,-	-,**15**,-	-,6,-	-,4,-	-,**16**,-	-,**16**,-	-,**16**,-	-,**16**,-	-,3,-	-,3,-	-,13,-	-,3,-	-,**5**,-
Correct	16/17	16/17	5/17	7/17	15/17	14/17	11/17	11/17	5/17	1/17	10/17	0/17	1.5,5,5 9,5
kCE	**15**	**15**	**15**	**15**	**16**	**16**	**16**	**16**	**5**	**5**	**2**	**1**	**12**

in every context in the experiments, but each CVAI suits a certain kind of data. This was the conclusion in (Arbelaitz et al. 2013), where a comparison of 30 different indices with 720 synthetic and 20 real datasets was made. In this study, Silhouette was nominated as the best index in general. Also the chosen clustering algorithm and initialization vary between experiments and affect the results.

The datasets tested in this study were all quite nicely distributed, including clear Gaussian clusters, with moderately low dimensions. This makes the job of the CVAIs a lot easier but in real life datas are seldom this straightforward. Nowadays, huge datasets can have tremendously many dimensions and often no clearly distinct groups. Therefore, the evaluation and comparison of the CVAIs should also be done with more complex a higher dimensional datasets.

6 Knowledge discovery from the activity of Finnish school children

This study consists of two parts. Both parts have been executed following the KDD process and using clustering as the data mining method. The goal of the first part is to find different activity profiles for the students on the weekly basis. From these profiles one can see, for example, at what times certain students have been the most active, does the activity behaviour differ between boys and girls etc. The second part deals with sedentary periods found in the data. The downsides, especially longer continuous periods, of sedentariness have been recognized even more lately as at the same time, especially amongst older children, the sedentariness has increased habitually (Hill et al. 2003). With clustering we can find a grouping based on the sedentary behaviour, try to find out conjunctive factors between the students in certain group, and by assessing these factors, try to find out ways to reduce the sedentariness in that cluster.

6.1 Recommendations and the Finnish schools on the move program

Many recommendations for the amount, kind, and intensity of physical activity at different ages have been made in different countries by the researchers (Organization 2010; Blair, LaMonte, and Nichaman 2004). In Finland, for instance, the UKK Institute has made recommendations for the age groups of 0-6, 7–18, 18–64, and over 65[1].

In Finland it is recommended (Varhaisvuosien fyysisen aktiivisuuden suositukset 2016), based on international recommendations, that children under six-year-old should have, at least, two hours of brisk activity a day. For children between the ages of seven and 18, 1-2 hours is recommended, so that seven-year-olds have two hours and then it evenly decreases so that 18-year-olds have an hour (Opetusministeriö & Nuori Suomi 2008). Out of this activity, half should be more intense. In addition, sedentary periods of longer than two hours should be avoided and screen time with entertainment media should be restricted to two hours.

1. http://www.ukkinstituutti.fi/ammattilaisille/terveysliikunnan-suositukset

The national Finnish Schools on the Move-programme[2], funded by the Ministry of Education and Culture and organised by the Board of Education, supports schools with promoting physical activity at school (Haapala et al. 2014). The goal is to encourage children to be physically active at school or immediately before or after (Tammelin, Laine, and Turpeinen 2012). The programme is being coordinated and evaluated by LIKES, Research Centre for Physical Activity and Health.

6.2 Study population and data collecting

The activity data was collected from 418 seventh-grade-students (age 13.7 ± 0.4), from nine different schools in the Jyväskylä school district in Finland. The collection was carried out during spring 2013 with 194 boys and 223 girls. Out of the students, 75 (18%) were on an exercise class, meaning that they had more physical education at school.

Physical activity was measured objectively using an accelerometer, ActiGraph GT3X (ActiGraph, Pensacola). The students were advised to wear the monitors on their waist for seven consecutive days, during waking hours, except when bathing, swimming, or participating in other water activities, since the accelerometer was not water-resistant. The time span for collected measurements varied a lot. Few students had worn the accelerometer for only a day whereas someone had recordings from 48 days.

Table 3. Distribution of the measurement days.

Days in Jan	Days in Feb	Days in Mar	Days in Apr	Days in May
763 (11.35%)	590 (8.77%)	3408 (50.68%)	1402 (20.85%)	561 (8.34%)

As metadata the gender of the student and information whether she or he was on an exercise class or not were considered. This background data was missing from one student. Also the amount of measurement days in certain month was calculated to see if the activity behaviour differed according to the time of the year. Table 3 shows the overall distribution of measurement days. In Section 6.6, we also used a categorical index indicating whether the student is of normal weight, overweight or obese, as metadata.

2. http://www.liikkuvakoulu.fi/

6.3 Preprocessing

The ActiGraph monitor stored activity counts for three axes once in a second (see Section 4.1.1). Since these counts are already band-pass filtered, no further filtering was done. In this study, only the counts from vertical axis were used for assessing the physical activity, because this direction provides the most significant deviator in the school environment. These counts per second were summed into counts per minute (CPM) for further transformation, in order to utilize the common limits for activity types defined using CPM (Evenson et al. 2008; Freedson, Pober, and Janz 2005; Trost et al. 2011). Periods of at least 30 minutes of zero counts were defined as nonwearing time as in (Syväoja et al. 2013).

6.4 Weekly physical activity of the students

6.4.1 Transformation

The activity counts from each student were transformed into a calendar form, resulting to a 7x24 table (see Figure 12), with rows representing the seven days of the week and columns the 24 hours of each day. The counts in certain hour were taken as the average of counts per minute over the 60 minutes. If the sensor had been worn for less than 30 minutes during that certain hour, it was marked as a missing value. This was done due to the assumptions in robust statistics – more than half of the measurements are needed for getting reliable results (Sprent and Smeeton 2016).

	hr 1	hr 2	hr 3	hr 4	hr 5	hr 6	hr 7	hr 8	hr 9	hr 10	hr 11	hr 12	hr 13	hr 14	hr 15	hr 16	hr 17	hr 18	hr 19	hr 20	hr 21	hr 22	hr 23	hr 24
Mon	-	-	-	-	-	-	-	-	942	266	127	205	267	550	451	826	658	1221	942	386	62	16	43	-
Tue	-	-	-	-	-	-	-	-	744	225	94	311	403	522	390	262	27	212	177	213	-	-	-	-
Wed	-	-	-	-	-	-	-	-	-	113	57	19	22	30	17	57	174	146	113	59	14	-	-	-
Thu	-	-	-	-	-	-	-	-	489	96	75	112	152	257	134	390	311	61	152	34	161	-	-	-
Fri	-	-	-	-	-	-	-	-	206	116	56	163	145	254	555	380	223	169	320	363	26	-	-	-
Sat	-	-	-	-	-	-	-	-	-	-	-	41	148	58	57	131	174	31	26	45	37	69	12	-
Sun	-	-	-	-	-	-	-	-	-	-	57	115	300	265	369	158	231	90	193	94	5	-	-	-

Figure 12. Average counts per minute for hours of the week from one student

These transformations were done separately for the entire time and for school time. To uncover more information about the activity during school, the time was divided into half hour periods, resulting into a 7x48 table. In Finland school classes normally start at quarter past and last for 45 minutes. Therefore, to find out if there was more activity during the

breaks between classes, the half hour periods were chosen to be from quarter past to quarter to and from quarter to, to quarter past. Again the average counts per minute over the 30 minute period were used for forming the table and if the sensor had been worn for less than 15 minutes, that certain half hour was marked as a missing value.

The transformed data, as well as the original data, had a lot of missing values due to the great amount of non-wearing time, especially during nights. The number of available observations for the hour and half-an-hour periods can be seen from Figures 13 and 14. Out of all 168 hours and 336 half-hours of the week, only the ones that had enough available measurements were chosen for the actual clustering (see Table 4). This selection was done manually so that with entire time 77% of the observations were available for the whole data and at least 40% for each separate variable and with school time 75% of observations were available for the whole data and at least 30% for each separate variable, as the robust spatial median is able to handle a great amount of missing values (Äyrämö 2006).

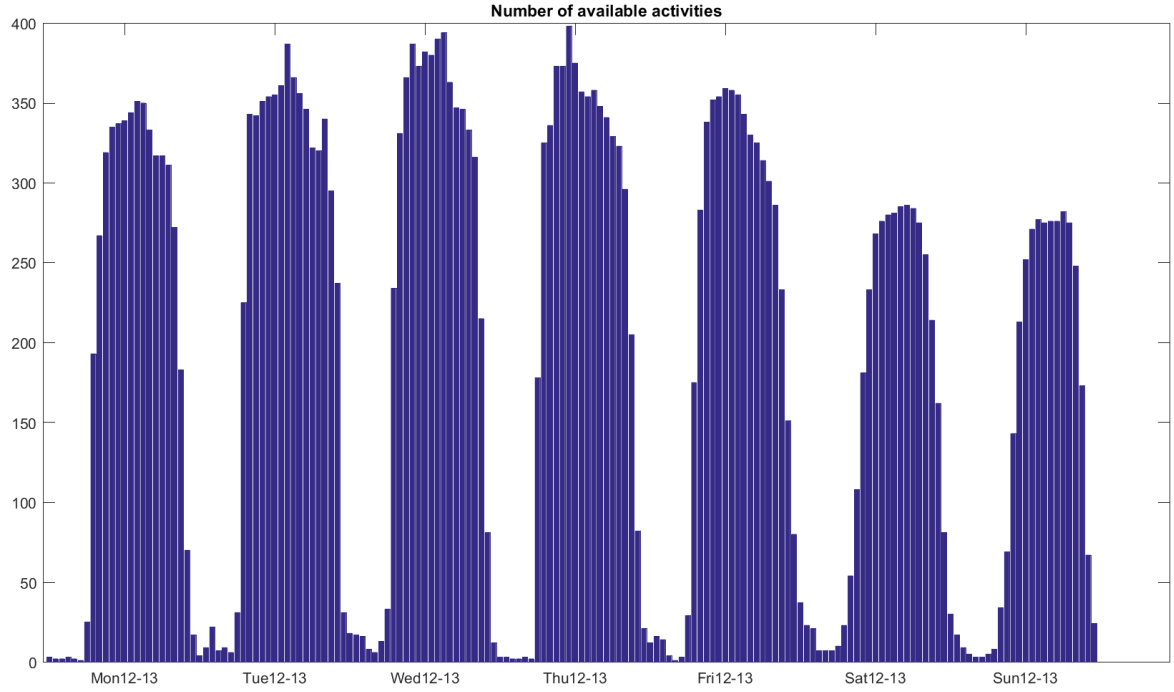

Figure 13. Available data for the one hour periods.

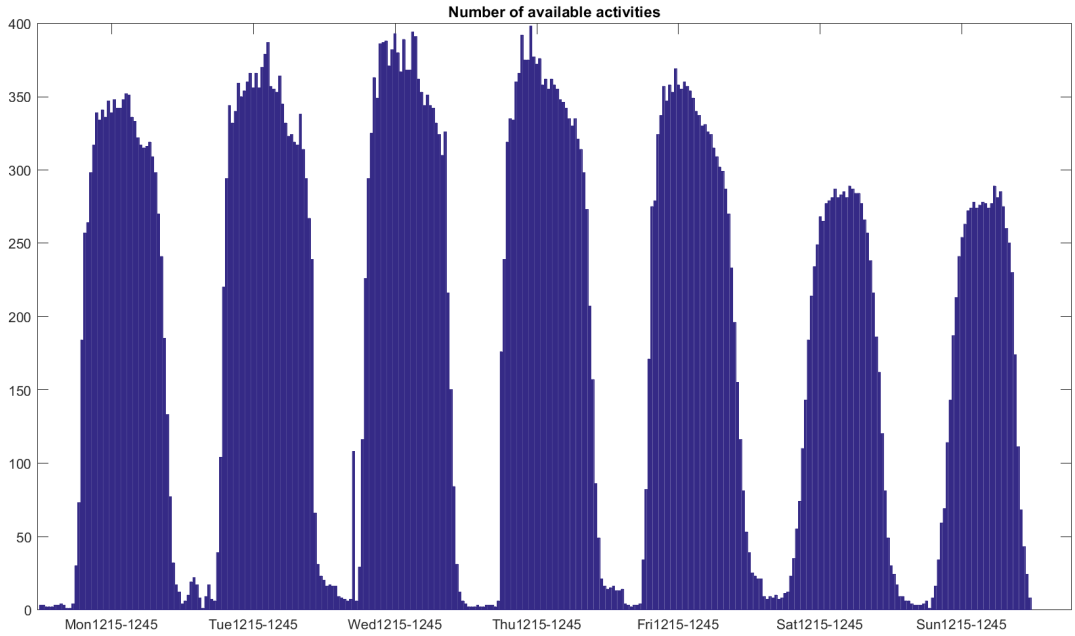

Figure 14. Available data for the half-an-hour periods.

Table 4. Variables chosen for clustering

	School time	Entire time
Hours		7-21 Mon-Fri 11-21 Sat-Sun
Half-hours	8.15-15.15 Mon-Thu 8.15-14.45 Fri	

6.4.2 Data mining

Data mining was done using robust clustering with the k-spatialmedians method (see Section 3.5). As a platform for this, MATLAB was used. Missing values were treated using the available data strategy, described in (Äyrämö 2006).

The number of clusters, K, was determined using multiple CVAIs – Ray Turi, both the original Davies Bouldin and its variation Davies Bouldin*, and robust Silhouette (Äyrämö 2006). Also the knee-point of the clustering error was considered, and the final number of clusters was defined according to majority voting – K suggested by most of the indeces was selected. All the CVAIs were tested with $K = 1 - 10$.

6.4.3 Results

To begin with, the cluster prototypes were ordered in an ascendant order with respect to the amount of overall activity (i.e., sum of the activity over all variables/times of the week). So the first cluster is always the one with the least activity whereas the last one has the most activity.

The mostly separating variables between clusters were determined utilizing non-parametric Kruskal-Wallis statistical test, as suggested in (Cord, Ambroise, and Cocquerez 2006; see also Saarela, Hämäläinen, and Kärkkäinen 2017). The results were then plotted with spatial median for each variable on the x-axis and the χ^2-value of Kruskal-Wallis on the y-axis (see, e.g., Figure 17 in Section 6.4.4). Then the most separating variables were visually observed and determined from the plots.

6.4.4 Entire time with one hour periods

Three groups, from now on referred as C1, C2 and C3, were found when clustering the 90 chosen hours considering the entire time. The prototypes for clusters can be seen in Figure 15 and as discussed before, they are ordered so that the students in C3 have the most overall activity and students in C1 have the least.

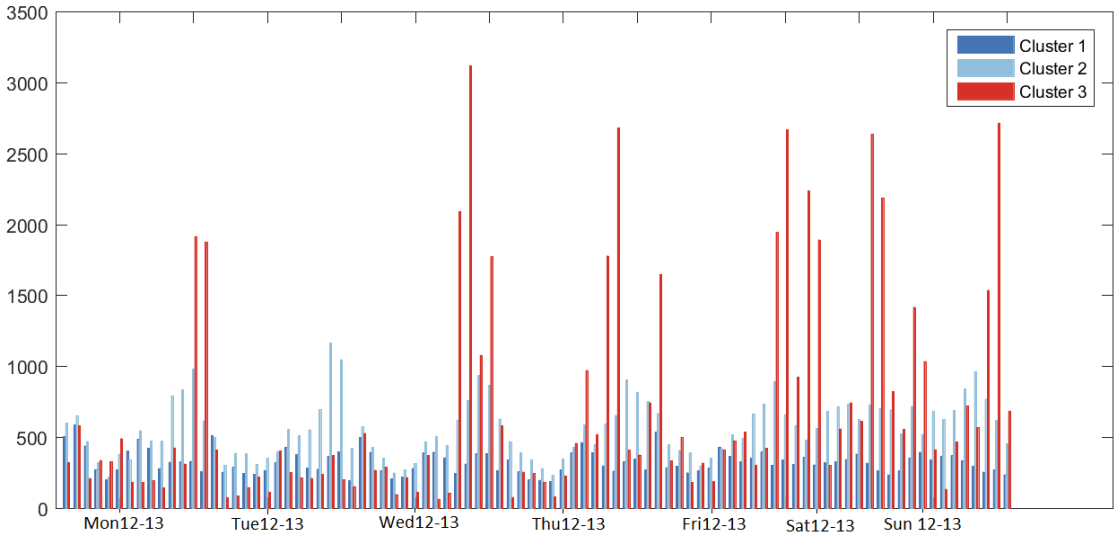

Figure 15. Cluster prototypes with one hour periods considering entire time. Cluster one is dark blue, cluster two is light blue and cluster three is red.

The most separating times between clusters, according to the Kruskal-Wallis χ^2-test, are on Sunday, Tuesday and Wednesday nights. These can be seen in more detail in Figure 16. As decribed before, these were determined from the plot in Figure 17 by visually observing.

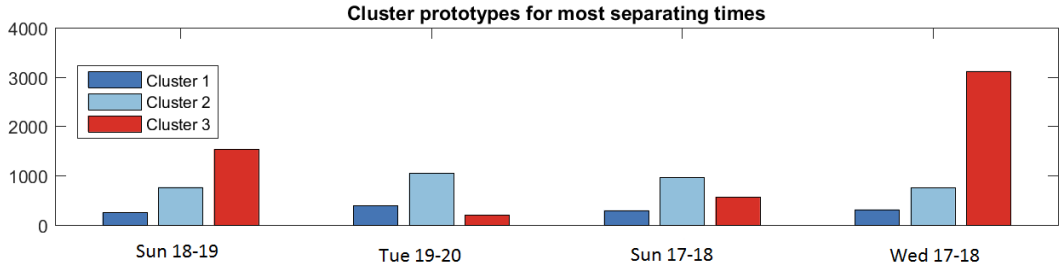

Figure 16. Cluster prototypes for the most separating times.

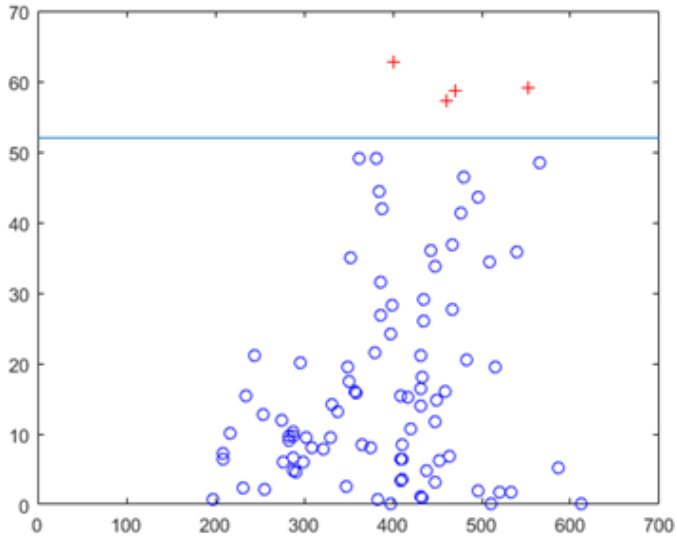

Figure 17. Spatial median for each variable on the x-axis and the χ^2-value of Kruskal-Wallis on the y-axis. The line was manually inserted and variables above it are the most separating.

Metadata for these clusters is described in Table 5. C1 is by far the biggest cluster (n=258), including more than half of the students. These students are clearly the least active throughout the week and concluding from the activity profile, the low activity is quite evenly distributed along the whole week, with no higher peaks. However, during weekdays some more activity can be seen during evenings compared to morning and afternoons. This cluster probably includes students that don't belong to any sports clubs or really do any sports at all. They spend majority of their time, over 96%, in sedentary or light activities. The distribution

of measurement days between different months is very similar to the original distribution of the months.

Table 5. Metadata for clusters with one hour periods considering entire time. Total activity is the sum of all activity over the week.

	Cluster 1	Cluster 2	Cluster 3
Size	258	138	22
Total activity	29522	50897	63462
Boys/Girls	41.1%/58.9%	52.2%/47.8%	72.7%/22.7%
Exercise class	12.8%	26.1%	18.2%
Part sedentary	66.0%	57.9%	63.6%
Part light	30.2%	34.8%	29.9%
Part moderate	2.7%	4.7%	3.8%
Part vigorous	1.0%	2.5%	2.6%
Days in Jan	10.2%	15.5%	0.0%
Days in Feb	6.0%	14.6%	3.4%
Days in Mar	59.5%	32.8%	63.7%
Days in Apr	18.9%	23.3%	26.2%
Days in May	5.4%	13.9%	6.6%

C2 is also a quite big cluster (n=138) having more than one-fourth of students on an exercise class. This cluster seems to have a lot more overall activity than C1. If only looking at the parts that the students in this cluster spend in different intensity activities, they would seem the most active ones. They have almost as much vigorous activity as students in C3 (2.5% vs. 2.6%) and the most moderate and light activity, when in turn the sedendary time is the smallest. However, their total activity is clearly smaller than in C3, due to not having as high peaks of activity. The activity is quite evenly distributed throughout the week (see Fig. 15), with some more activity in the evenings and on the weekends, so on their free time. In this cluster, the measurement days are distributed more evenly between different months, with respect to the original distribution of months. In practice, less measurements are taken in March and more during the other measurement months.

C3 is a small cluster (n=22) that has far more boys (72.7%) than girls in it. The activity profile shows that while during daytime at school these students are often the most sedentary, in the evenings they have overwhelmingly more intense activity compared to the other two clusters. This cluster probably includes students that have sport practices after school and/or on the weekends. However, outside these intense activity times they are quite sedentary, spending over 93% of their time in sedentary and light activities. Interestingly, the most active cluster C3 is not characterized by the students belonging to the exercise class. Majority of the measurements (90%) in this cluster are taken during March and April. In the original distribution, March and April cover less than 72% of the measurement days.

6.4.5 School time with half-an-hour periods

When half-an-hour periods from the school time were considered, four clusters, C1, C2, C3, and C4, were found. The prototypes, again ordered based on the overall activity, can be seen in Figure 18.

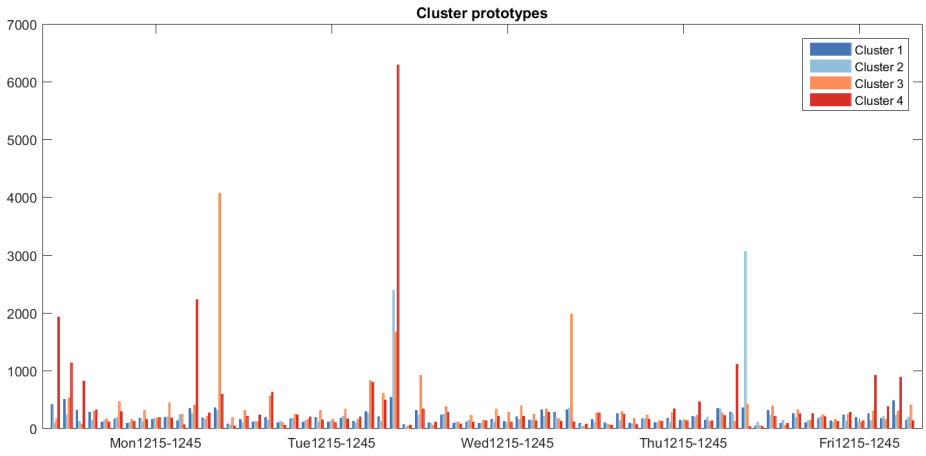

Figure 18. Cluster prototypes with half hour periods considering school time. Cluster one is dark blue, cluster two is light blue, cluster three is orange and cluster four is red.

The most separating times between these clusters can be seen in Figure 19. These include Monday and Tuesday mornings (9:45 - 10:15) and afternoons (14:45 - 15:15) and Tuesday 11:45-12:15. These were again determined based on the plot in Figure 20.

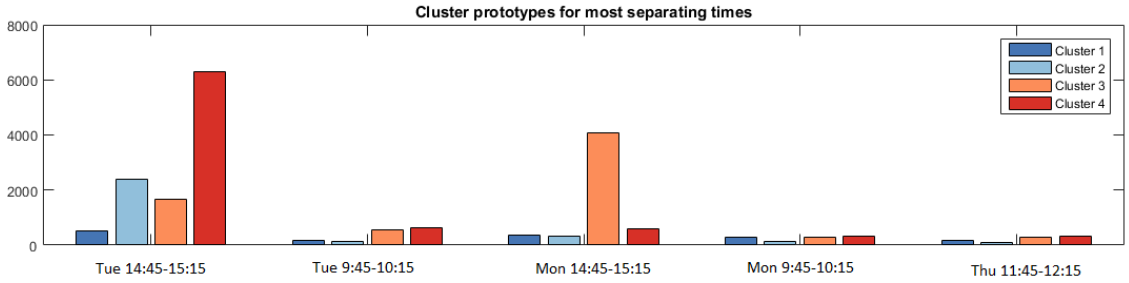

Figure 19. Cluster prototypes for the most separating times.

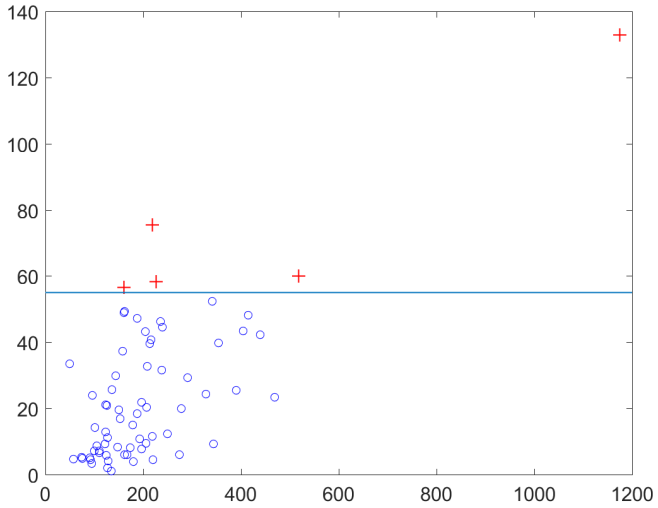

Figure 20. Spatial median for each variable on the x-axis and the χ^2-value of Kruskal-Wallis on the y-axis. The line was manually inserted and variables above it are the most separating.

Metadata for these clusters can be seen in Table 6. C1 is the cluster with the least overall activity, having 137 students in it. The activity in this cluster is quite evenly distributed between all school days. There are two peaks, on Tuesday and Thursday afternoons, that are probably physical education classes at school. Otherwise, they are quite passive during school and have the most sedentary time. Therefore, it is no surprise that this cluster has very little exercise class students, since they should be having more activity during school. This cluster had almost no measurements in January and February, but clearly more in April and May.

Table 6. Metadata for clusters with half-hour periods considering school time. Total activity is the sum of all activity over the week.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Size	137	51	198	32
Total activity	13716	15888	24787	27826
Boys/Girls	46.7%/52.6%	45.1%/54.9%	44.9%/55.1%	56.3%/43.7%
Exercise class	2.0%	26.1%	26.3%	3.1%
Part sedentary	68.1%	64.3%	65.5%	67.2%
Part light	28.9%	31.1%	30.5%	28.8%
Part moderate	2.3%	3.3%	2.9%	2.8%
Part vigorous	0.6%	1.3%	1.0%	1.2%
Days in Jan	0.0%	15.7%	19.6%	3.3%
Days in Feb	1.1%	0.0%	16.9%	0.0%
Days in Mar	51.8%	60.5%	44.8%	77.3%
Days in Apr	26.4%	23.8%	16.5%	19.4%
Days in May	20.6%	0.0%	2.2%	0.0%

C2 is a smaller cluster (n=51), with a little bit more exercise class students. The activity in this cluster is very evenly distributed and it is the only cluster without any clear activity peaks. This is again a surprise, because the students on exercise classes should be having more physical education at school.

C3 is the biggest cluster (n=198) that also has the most exercise class students. Similarly to C1, the activity is quite evenly distributed, with a few peaks. The highest peak is on Monday afternoon and a few lower peaks are on Tuesday and Wednesday afternoons. These are again probably the times when the students have physical education classes at school.

The overall most active cluster, C4, is a small cluster (n=32) with not many exercise class students. They have the most activity peaks during schooldays but most of these peaks are a little bit lower than the peaks of other clusters. One of the peaks is also very high. These lower peaks could be either physical education or breaks during lessons. The high peak is

probably a physical education class. A lot of the measurements in this cluster were taken in March (77%) and most of them during March and April (97%).

6.4.6 Conclusion

With clustering we obtained different groupings for the students based on their activity behaviour. In the results, the metadata between clusters was not significantly different. Boys and girls were generally evenly distributed between the clusters, while just a few clusters had clearly more students from the exercise classes. The measurement days from different months also distributed quite evenly to the clusters, so no clear conclusions could be made based on them. Therefore, with clustering we found a new grouping for the students, that was not explained by the metadata behind them. This approach has potential in recognizing certain kind of activity behaviour and this way making more suitable and personalized actions when trying to encourage the students to be more active.

We found that the overall behaviour in the study population was quite sedentary, which supports many previous findings (Salmon et al. 2005; Tammelin et al. 2007; Pate et al. 2011). Interestingly, the majority of exercise class students were not in the most active clusters, not even when just school time was considered, even thought they are supposed to have more physical education at school.

As the students in this study are from the same region, schools and even classes, the activity behaviour and the grouping are, at least partly, explained by their common physical education classes and sport exercises. However, in a case where the subjects do not have this much common background, this approach has even more potential to discover novel and interesting information about activity behaviour.

6.5 Sedentary behaviour of the students

6.5.1 Transformation

The sedentary behaviour of the students was examined during their free time, because during school hours a lot of sedentary time exists due to the classes. Times between 16 and 23 were

considered from all days, during both weekend and week. The counts were already summed to CPM during preprocessing and a limit of 100 CPM was chosen for sedentary activity (Evenson et al. 2008). In Figure 21, the sedentariness can be seen as CPM less than 100. If there is even one CPM that exceeds the limit, the sedentary period ends. The non-wearing time is defined as 30-minutes of zero counts, and not taken into consideration when observing sedentary behaviour.

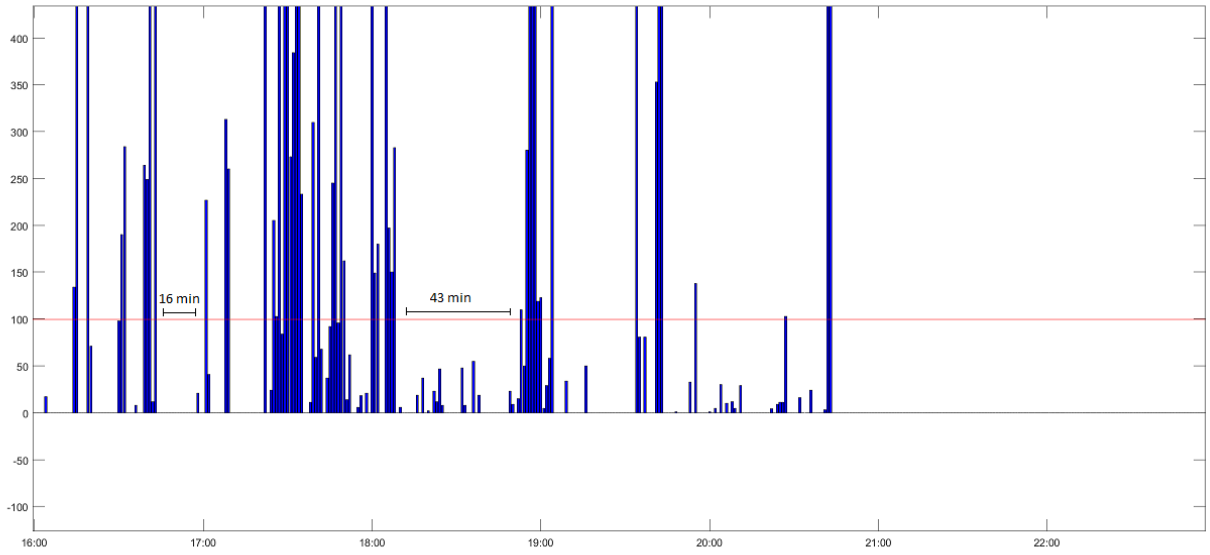

Figure 21. Part of activity counts form a student. CPM less than 100 is considered sedentary time.

A table of sedentary periods was formed so that the amount of sedentary periods of length n minutes is given in the corresponding column n (See Table 7). Each row represent a day from a student and the two last columns include metadata, i.e., the number of the student and the day of the week for that row.

After the table was formed columns and rows that did not have any observations were dropped out. So if none of the students had any sedentary periods of, e.g., length 200 minutes, column 200 was dropped and if certain day from a student had no sedentary periods on any length, the sensor was probably not worn on that day and the row was dropped. When first considering the whole data, i.e., taking all available measurements instead of the seven

Table 7. Example table of sedentary periods of the students.

1	2	3	4	...	360	Student	Weekday
5	8	18	6	...	0	1	1
13	12	3	4	...	0	1	2
7	24	14	13	...	0	1	3
15	8	3	24	...	0	1	4
21	10	23	4	...	0	1	5
9	8	13	14	...	0	1	6
4	9	5	16	...	0	2	3
⋮	⋮	⋮	⋮	...	⋮	⋮	⋮
15	8	3	4	...	0	417	7
16	22	15	4	...	0	418	2
27	8	3	18	...	0	418	3
5	25	13	0	...	0	418	4
23	8	3	8	...	0	418	5

official measurement days, this led to a data with 3836 observations and 141 variables, so approximately nine days from each student (transformation 1).

Next, exactly the same transformation was done so that instead of considering the all available data, only the seven official measurement days were considered. This resulted into a data with 2888 observations and 116 variables (transformation 2).

Different scaling methods were also tested to see how they affect the data mining. The distribution of the data was very skewed in both transformations, because there were always more short sedentary periods than the longer ones. Therefore, logarithmic scaling was tested in order to smooth the distribution. Min-max scaling to range $[-1, 1]$ was also done to even the differences in general and because it had been found to work well with our test dataset when evaluating CVAIs in Chapter 5.

In (Kim et al. 2015) only sedentary periods longer than 10 minutes were associated with increased health risks. As we were also interested in these longer, harmful sedentary periods and did not want the shorter ones to rule too much in data mining, we also tested leaving the first ten columns, i.e., sedentary periods of length $1-10$ minutes, out.

From the transformation 2, three new transformations were formed. The purpose was to form a week and weekend observation for each student. The weekday observation was obtained by combining all the weekday observations (mon-fri) from the student and weekend observation by combining the weekend observations (sat-sun).

- The weekday observation for a student was taken as mean of all weekday observations from him/her, weekend days similarly. This resulted into 799 observations and 116 variables (transformation 2.1).
- The weekday observation for a student is taken as maximum of all weekday observations from him/her, weekend days similarly. This resulted into 799 observations and 116 variables (transformation 2.2).
- Similar idea as in the above cases, but first min-max scaling the data to the range of $[-1,1]$ and then taking the mean over students weekdays and weekends. From this data, sedentary periods shorter than 10 minutes were dropped out and a 799×106 dataset was obtained. To this end, principal component analysis (see Section 2.3.1) was done and by choosing components that explain more than 95% of the total variance, a dataset of size 799×57 was obtained (transformation 2.3).

So altogether five different transformed representations were formed. Notice that when combining the week and weekend observations the use of maximum value resulted into a more informative representation with reasonable sized integers, while the use of mean resulted into much smaller values in general. However, the mean more accurately represents the overall behaviour of a student, while taking the maximum of the sedentary periods probably makes the student seem more passive than he/she actually has been.

6.5.2 Data mining

Data mining of sedentary periods was done using MATLAB's k-means++ algorithm. The number of clusters was defined with a set of evaluated CVAIs. This set consisted of the seven CVAIs that worked best in our tests with multiple datasets (See Chapter 5), including Silhouette, Calinski Harabasz, Davies Bouldin, PBM, Wemmert Gançarski, Generalized Dunn, and Ray Turi. Own implementations of these indices were made in MATLAB, some slightly differing from the implementation in packages used in testing. Also for CVAIs that proposed the number of clusters at the maximum index value, an inverse was taken so that the results between CVAIs would be easier to compare. The CVAIs were tested with number of clusters $K = 2 - 20$.

For transformation 1 the number of clusters was chosen as 5 based on the values of CVAIs. Those CVAIs that best supported this $K = 5$ can be seen in Figure 22. The Davies Bouldin index suggests $K = 20$ as first, but a clear knee-point can be seen in the curve at $K = 5$.

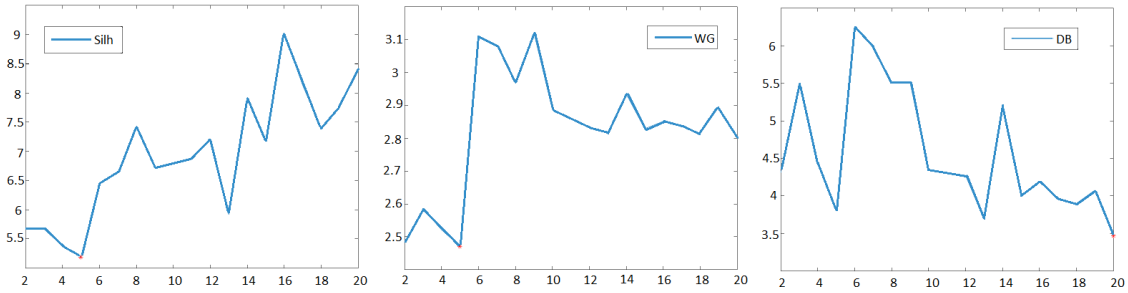

Figure 22. CVAIs for transformation 1 that support the number of clusters being 5. In y-axis the value for CVAI and in x-axis the number of clusters, K. From left to right the CVAIs are Silhouette, Wemmert Gançarski, and Davies Bouldin. The proposed number is the minimum index value.

When these five clusters were formed with k-mean clustering, two of them were significantly larger. For these two larger clusters the same process was hierarchically performed as in (Wartiainen and Kärkkäinen 2015). For the first bigger cluster the CVAIs suggested that six subclusters could be found and for the other five. See the result of clustering in Figure 23. So when the k-means was hierarchically done to transformation 1, altogether 16 clusters were obtained.

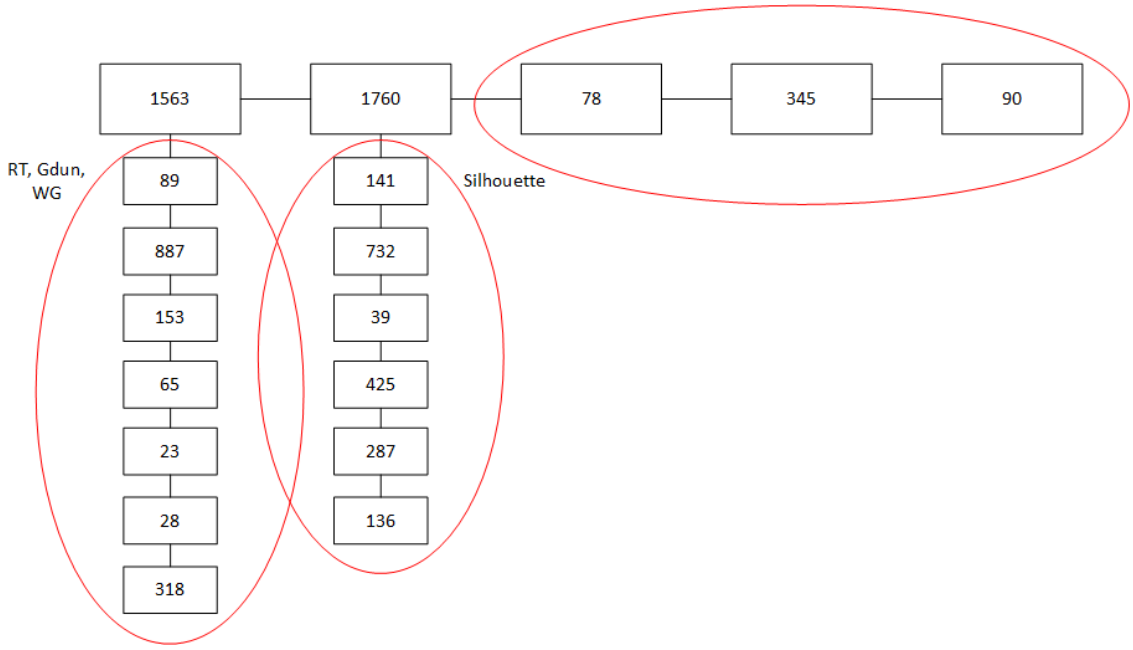

Figure 23. The results from hierarchical prototype-based clustering. First five cluster were formed and then the two largest ones were clustered again to find subgroups (Wartiainen and Kärkkäinen 2015; Saarela and Kärkkäinen 2015). The sizes of clusters are marked on the figure as well as the indices proposing the number of subclusters found.

No data mining was done for transformation 2, but only for further processed transformations 2.1, 2.2, and 2.3. With transformations 2.1 and 2.2, the CVAIs all proposed a different number of clusters and the number of clusters could not be defined even when investigating the plotted curves of their values. This happened also when logarithmic transformation, min-max scaling, and dropping out the first ten columns were tested. So no results were obtained from these transformations.

For transformation 2.3, the number of clusters, K, was chosen as ten based on the values of CVAIs. Those CVAIs that best supported $K = 10$ are illustrated in Figure 24. The Ray Turi index proposed eight for K but as can be seen from the plotted values, $K = 10$ is also very close to the minimum.

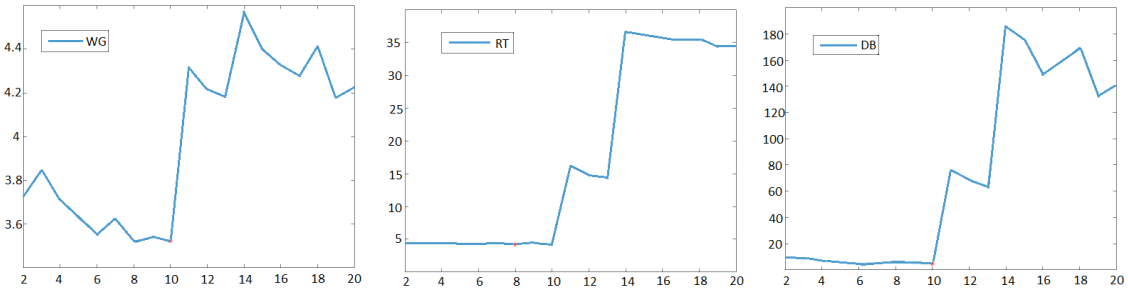

Figure 24. CVAIs for transformation 2.3 that support the number of clusters being 10. In y-axis the value for CVAI and in x-axis the number of clusters, K. From left to right the CVAIs are Wemmert Gançarski, Ray Turi and Davies Bouldin. The proposed number is the minimum index value.

6.5.3 Results

The cluster prototypes are ordered from the most passive to the least passive. So the first cluster has the largest overall amount of sedentary time and the last has the smallest amount. The most separating variables between the clusters were simply defined with the total difference of sedentary time between the prototypes, generalizing the approach proposed in (Saarela and Kärkkäinen 2015). This was done so that the prototypes for each variable were first ordered into ascending order and then the difference between the ordered values was computed and summed up. This created the total difference for that variable, and when these were ordered into descending order, the most separating variables were found.

6.5.4 Transformation 1

With transformation 1 each day from a student is an observation and 16 clusters were found when hierarchically applying the k-means method. The metadata for these clusters can be seen in Table 8. The column "Students" indicates from how many different students the observations come, column ">60" is the percentage of sedentary time that comes from periods longer than 60 minutes, and column "<10" is the percentage that comes from short, less than ten minutes, periods. The prototypes for the ten most separating variables can be seen in Figure 25 and the parts of certain weekdays in the clusters in Figure 26.

When the separating variable is mostly present in one cluster (high peaks in the figure), it can

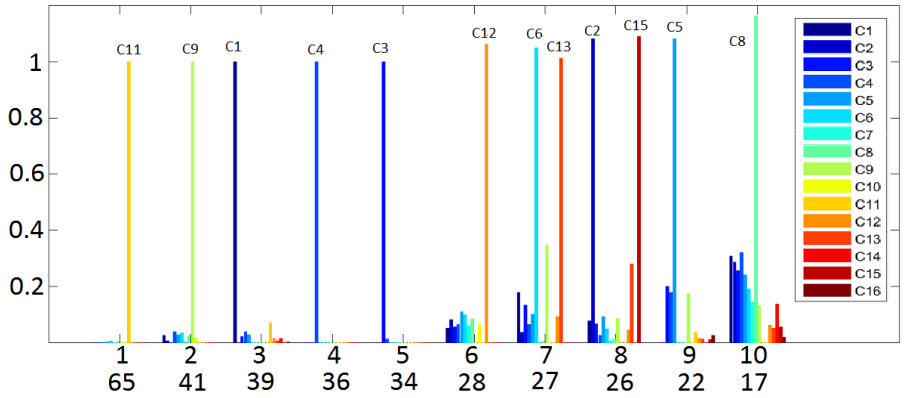

Figure 25. Transformation 1, 10 most separating variables. For example, the most separating variable is a sedentary period of length 65 minutes, mostly present in cluster 11. This variable can be considered characterizing this cluster, since it is the one mostly separating it from the others.

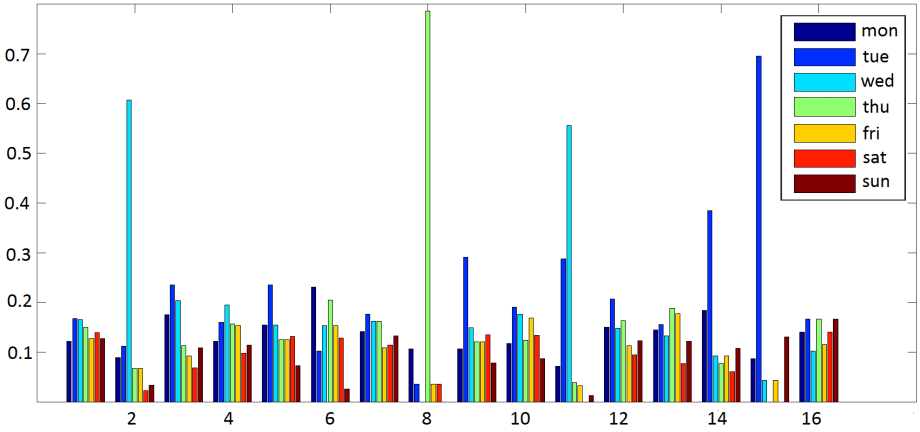

Figure 26. Transformation 1, portion of a weekday in a cluster.

be thought as a characteristic variable for that cluster. That variable is the one separating this cluster from the others, as it is not as present in other clusters. Most of the metadata is quite evenly distributed between the clusters, and does not provide explanation of the clusters. However, there were some exceptions. Cluster 15 has a lot (70%) boys, clusters 3,8, and 11 in turn have a lot of their sedentary time coming from periods longer than 60 minutes, which is a concern. Some clusters also had significantly more certain weekdays as seen in Figure 26. For example, 61% of the days in C2 and 79% in C8 are Wednesdays, while 70% in C15 are Tuesdays.

Table 8. Metadata for clusters obtained with transformation 1

	Size	Students	Boys	Ex Class	>60	<10
Whole data	3836	418	46%	19%	6%	44%
C1	345	235	41%	14%	3%	41%
C2	89	81	49%	22%	8%	20%
C3	887	347	52%	12%	33%	22%
C4	287	210	43%	17%	3%	44%
C5	136	110	38%	13%	1%	45%
C6	39	39	36%	21%	4%	40%
C7	732	338	46%	18%	3%	50%
C8	28	26	32%	7%	51%	13%
C9	141	123	43%	11%	3%	45%
C10	425	235	45%	14%	1%	67%
C11	153	127	48%	23%	25%	13%
C12	318	211	48%	23%	9%	36%
C13	90	83	46%	9%	4%	35%
C14	65	61	43%	20%	16%	18%
C15	23	23	70%	30%	12%	7%
C16	78	72	37%	17%	2%	35%

The size of the cluster indicated the amount for days assigned to that cluster and as stated before, "students" tells from how many different students these days come from. Looking at these numbers, we can see that the days from certain student did not all go into the same cluster but rather divided between many clusters. Also the bigger the cluster the more students there were that had a day in it. It means that a lot of the students have this kind of similar days. For example, more than 80% of the students have a day in C3 and 33% of the sedentary time in it comes from periods longer than 60 minutes. By taking a closer look at this cluster, ways to intervene these periods might be found.

As there were multiple days from each student and they were scattered between the clusters and did not go into the same one, they can be deduced to be very different between each

other. This indicates that when doing any kind of studies about the activity of children, multiple days should always be considered in the measuring process.

6.5.5 Transformation 2.3

With transformation 2.3 the interest was on students week vs. weekend observations and ten clusters were found. The metadata for these clusters can be seen in Table 9. The prototypes for the ten most separating variables can be seen in Figure 27. Here again, when observing the size of the cluster and from how many different students these days come from, we can see that the observations from one student mostly did not go into the same cluster.

Table 9. Metadata for clusters obtained with transformation 2.3

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Size	61	46	24	51	237	28	222	24	91	15
Students	58	45	24	51	179	28	192	24	80	15
Boys (46%)	43%	37%	29%	39%	57%	43%	40%	46%	48%	47%
Exercise class	8%	20%	38%	16%	24%	4%	16%	8%	16%	7%
Weekdays (51%)	49%	80%	4%	33%	54%	32%	67%	17%	40%	13%
Overweight (11%)	16%	13%	25%	10%	12%	7%	6%	17%	11%	7%
Obese (2.5%)	1.6%	0%	0%	5.9%	2.1%	0%	3.6%	8.3%	2.2%	0%

The metadata does not differ much between the clusters. It is again interesting how the exercise class students seem to be just as sedentary as all the other students. Cluster 2 has clearly the most weekday observations (80%), while cluster 3 has almost only weekend observations (96%). Cluster 3 also has some other interesting characteristics – it has the most exercise class students but also the most overweight students. It also has clearly the most girls in it.

As the observations from each student were scattered between clusters, we can again conclude that the sedentary behaviour of a student differs between week and weekend.

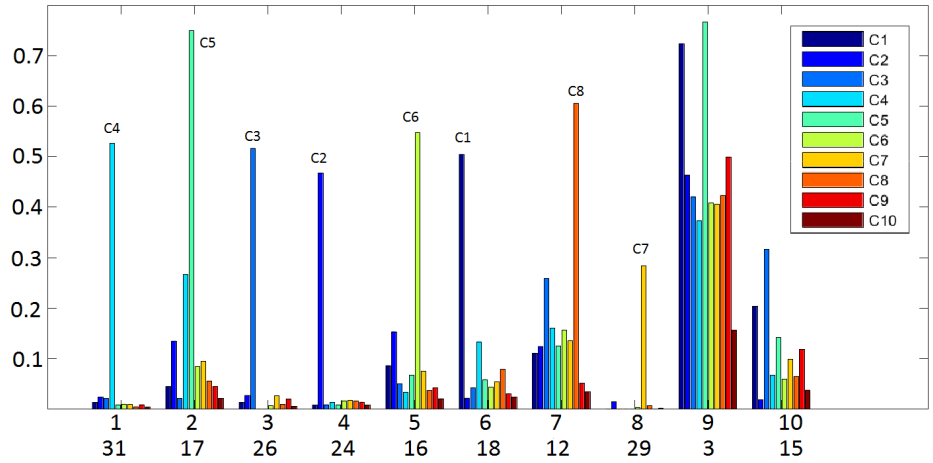

Figure 27. Transformation 2.3, 10 most separating variables. For example, the most separating variable is a sedentary period of length 31 minutes, mostly present in cluster 4. This variable can be considered characterizing this cluster, since it is the one mostly separating it from the others.

6.6 Sedentary behaviour of the students 2

6.6.1 Transformation

In the previous transformation, all sedentary periods with the accuracy of one minute were considered. However, there is no significant difference in, for example, sedentary periods of 45 or 46 minutes. Therefore it might be more reasonable to investigate combined periods. Also instead of clustering the week and weekend observations from students, we might want to cluster just the students, so a combination of these observation is needed.

Beginning with the same representation as in the previous section (see Table 7 in Section 6.5.1), the week and weekend observations were again formed by taking the mean of student's week and weekend observations. This resulted into similar form as transformation 2.1 in the previous Section 6.5.1, with 799 observations and 116 variables. Next the sedentary periods were combined into ten different categories – 10-15, 16-20, 21-25, 26-30, 31-35, 36-40, 41-50, 51-60, 61-90 and 91-181 minutes. Sedentary periods shorter than ten minutes were again ignored. So now there were ten variables instead of 116.

The combination of student's week and weekend observations was done in a way that these

would still be separable and the comparison of sedentary behaviour during week and weekend was possible. So, the weekend row was joined together with the week row, resulting into 20 variables for each student. The students that were missing either of these observations, for example, due to not wearing the sensor during weekend, were ignored. This resulted into 386 observations (students).

6.6.2 Data Mining

The data mining was again done using MATLAB's k-means++ algorithm. This time, however, the variables were median values, instead of mean, so the city-block distance was used. This corresponds to the k-medians method, i.e., for $p = q = 1$ in Section 3.4.

The number of clusters was defined with our own implementations of the CVAIs found best in our comparisons. This set of indices consists of Silhouette, Davies Bouldin, PBM, Wemmert Gançarski, Generalized Dunn, and Ray Turi in their previously presented forms (see Section 3.6), also run with $p = q = 1$. For the CVAIs that proposed the number of clusters at the maximum index value, an inverse was taken so that the results between CVAIs would be easier to compare. The CVAIs were tested with number of clusters $K = 2 - 20$.

6.6.3 Results

The CVAIs indicated that the data contains three clusters. Metadata for these clusters can be seen in Table 10. In Figure 28, the difference between the whole data median and the prototypes can be seen. According to the Kruskal-Wallis test, the three groups were significantly different ($p < 0.01$) with all the variables, i.e., sedentary periods.

Table 10. Metadata for clusters.

	Size	Boys	Ex Class	Overweight	Obese
C1	210	51%	22%	11%	3%
C2	145	37%	12%	10%	3%
C3	31	52%	26%	13%	0%

The first cluster, C1, is the biggest cluster (n=210), including more than half of the students.

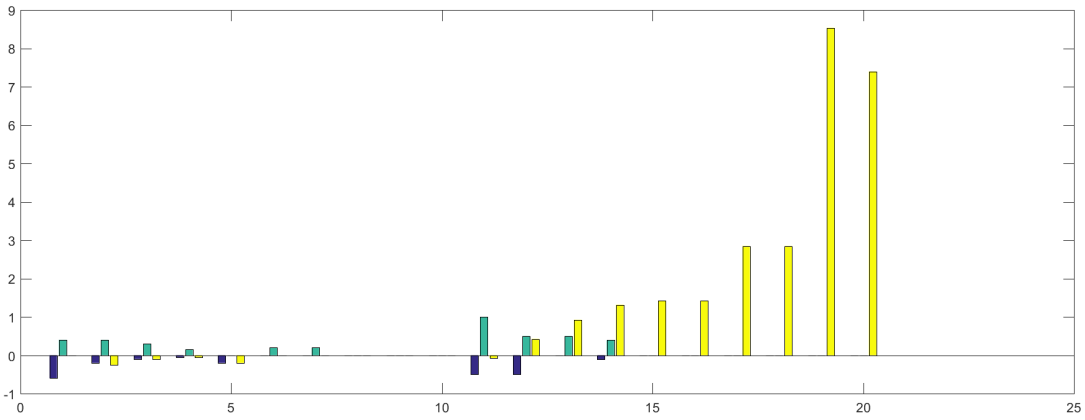

Figure 28. The difference between prototypes and the whole data median. C1 is blue, C2 is green, and C3 is yellow.

It is a cluster where the students have less sedentary time compared to the whole population and also they have almost no longer periods. Their behaviour is very similar through the week and weekend days.

The second cluster, C2, is quite similar to C1, mainly just shorter sedentary periods. However, they have a little bit more sedentary time when compared to the whole population and some longer periods during the weekdays.

The most interesting group is the third cluster, C3, that is a small cluster (n=31) with a very different sedentary behaviour than the other clusters. During the weekdays these students have less sedentary time compared to the whole population, but during weekends they have significantly more and longer sedentary periods than the population in general. Especially the amount of these long sedentary periods is a concern and this behaviour should be intervened. It is again interesting how this most sedentary cluster has the largest proportion of exercise class students.

Again, with clustering we have achieved a new grouping, which could not have been achieved utilizing the metadata, since it is fairly evenly distributed between the clusters. Also, we have found a very interesting group of students, C3, whose sedentary behaviour and reasons behind it should be further researched. Furthermore, if some common factors for this behaviour are found they can lead to novel and very useful information that can be utilized by domain

experts when making general recommendations or trying to come up with ways to intervene this kind of behaviour.

6.7 Summary

In this chapter, we applied between-method triangulation to students activity data. In general, triangulation is an approach where the same research objective is studied with either different datas, theories, researchers, or, as in between-method triangulation, analysis methods (Denzin 1970). We investigated both the active and sedentary time of the students, both with many different transformations and amany clustering and cluster validation approaches.

To asses the activity behaviour on a weekly basis, we first formed a calendar form representation in Section 6.4. The activity was studied in one hour and half-an-hour periods. With the former, three groups were found and with the latter four groups. We found that the general activity behaviour in the whole population was quite sedentary, most of the time was spent in sedentary or light activities.

Next we studied the sedentary behaviour of the students, forming multiple different representations of it in Section 6.5. First, when having each day from each student as an observation and using hierarchical prototype-based clustering (Wartiainen and Kärkkäinen 2015) we found 16 clusters. Second, the days were combined so that we obtained a week and a weekend observation for each student by taking the mean over their weekdays and weekend days. This way we found ten clusters. As in both of these cases the observations from a certain student did not go into the same cluster, but rather scattered along them, we can conclude that the daily activity of a student is different on each day. Therefore, when measuring the activity, multiple days should always be considered and included in the measurement period.

In Section 6.6, we formed just one observation for each student, including information from both the week and weekend. Three clusters were found, and one them was particularly interesting due to their sedentary behaviour during weekends. Even though during weekdays the behaviour was quite similar among the whole population, on the weekends cluster 3 contained overwhelmingly more and longer sedentary periods than the others. This group should be further investigated in order to find possible reasons behind this behaviour and

trying to intervene it.

In all the results, the metadata of the students did not explain the found clusters, so a set of novel profiles was obtained. Also, surprisingly, the students on an exercise class did not seem to be more active, but on the contrary, the least active and most sedentary clusters often had the largest proportion of those students. However, this matter should be further researched before any general conclusions can be made.

Also common for all the results, was that they included different kind of activity and sedentary profiles. For example, some students had a moderate amount of total activity that was divided very evenly over the week, whereas some students had some very high peaks of activity and otherwise they spent their time quite sedentary. Both of these groups could be approached in a different way in order to optimize their activity behaviour and its health benefits. For the first group, some more intense activity could be organized and encouraged while the students in the second group could benefit from some more light activity in addition to the higher peaks. This could also have potential in preventing injuries, as their activity could become more versatile instead of possibly just certain sport specific training.

Also the sedentary profiles have a large potential in providing a novel and useful way to intervene the sedentariness. A new way of grouping was obtained solely based on the sedentary behaviour that enables planning more personalized ways to affect the students.

7 Conclusion

The importance of physical activity has become an undisputed fact, but at the same time the amount of it has continuously decreased, while the sedentary time has increased (Owen et al. 2010). This, among other things, has lead to physical activity being a very current and interesting study field. The advancing technical abilities to measure different components of physical activity have lead to large amount of available data and many new applications

This thesis is focused on the KDD process and its utilization with physical activity data and applications. KDD's potential to utilize very large datasets, that are becoming more common in the field of physical activity, makes it a natural framework for this domain. With multiple transformations of the data and between-method triangulation we obtained many results that all both brought some novel information while also supporting each others findings.

Throughout the empirical part of the thesis, we followed the KDD process, using clustering as the data mining method. To validate the number of clusters present in each transformation of the student data, many CVAIs were used. These were chosen based on our tests with a extensive set of different indices and their implementations for the k-means clustering in MATLAB and R. These results have also been published in (Jauhiainen and Kärkkäinen 2017) and the best performing CVAIs were further generalized in this work and used also with the k-means and k-spatialmedians methods.

As we tried to represent the data with many versatile transformations, we ended with different measures and sparsities of data. It was very important to choose the data mining methods, missing data handling strategies, and distance measures case dependently. When the data was very sparse, having a lot of missing values, we used robust clustering – spatial median with available data strategy. With the k-means method, the used distance measure had to be chosen carefully. With discrete variables, e.g., median, we used the cityblock distance, while with continuous, e.g., mean, we used the euclidean distance. The used distance had to be considered also when validating the number of clusters with the CVAIs.

Our first approach was to form a calendar form of the activity of students and with this transformation we were able to access the division of activity along the week. We investigated

different time slots, hour and half-an-hour, considering both the entire time and only the time at school. We found a new grouping based on only the amount of activity at certain times of the week and this has the potential to provide a new way to affect the students activity behaviour in a more personalized way. For example, some students could benefit considerably from adding a few more intense activities to their week, while other could benefit from having more light activity throughout the week.

Second approach was to, by clustering, find different profiles based on the sedentary behaviour of the students. The sedentariness of children has become a concern (Tammelin et al. 2007) and the aim was to, through clustering, find different sedentary profiles and some habitual factors behind their sedentary periods. Many different transformations and measures were tested and finally novel representations and groupings were found that have a lot of potential in helping to recognize harmful sedentary behavior of students. Particularly, we found a group of students that are extremely sedentary during their weekends. In further research, if some reasons behind this sedentariness are found, they can be very useful in recognizing and preventing similar behaviour in other students.

With all the approaches and transformations, the metadata of the students did not explain the grouping and therefore with clustering, new and useful groups were found. This proves that clustering has a huge potential in offering new ways to access the activity and sedentary behaviour of the students and finding more personalized and effective ways to affect in it. Also, surprisingly, the students on exercise classes did not stand out as the most active or least passive in any of our results. This matter might need some further research, so that their behaviour and its health benefits could also be optimized more. Another important finding was that the days of a student were very different from each other throughout the week and therefore, when measuring activity of students in the future, multiple days should always be covered. So to summarize and answer the research question:

Unsupervised data mining, clustering in this case, can enable the finding of novel and very useful information from students activity data. Moreover, the use of different approaches and transformations provides also multiple different results. Finally, utilization of all this new information should happen together with domain experts.

Bibliography

Achten, Juul, and Asker E Jeukendrup. 2003. “Heart rate monitoring”. *Sports medicine* 33 (7): 517–538.

Aggarwal, Charu C, and Jiawei Han. 2014. *Frequent pattern mining.* Springer.

Agrawal, Rakesh, Ramakrishnan Srikant, et al. 1994. “Fast algorithms for mining association rules”. In *Proc. 20th int. conf. very large data bases, VLDB,* 1215:487–499.

Ainsworth, Barbara E, William L Haskell, Melicia C Whitt, Melinda L Irwin, Ann M Swartz, Scott J Strath, WILLIAM L O Brien, David R Bassett, Kathryn H Schmitz, Patricia O Emplaincourt, et al. 2000. “Compendium of physical activities: an update of activity codes and MET intensities”. *Medicine and science in sports and exercise* 32 (9; SUPP/1): S498–S504.

Aldenderfer, Mark S, and Roger K Blashfield. 1984. “Cluster analysis. Sage University paper series on quantitative applications in the social sciences 07-044”.

Arbelaitz, Olatz, Ibai Gurrutxaga, Javier Muguerza, Jesús M Pérez, and Inigo Perona. 2013. “An extensive comparative study of cluster validity indices”. *Pattern Recognition* 46 (1): 243–256.

Arbia, Dhafer Ben, Muhammad Mahtab Alam, Rabah Attia, and Elyes Ben Hamida. 2015. “Data dissemination strategies for emerging wireless body-to-body networks based internet of humans”. In *Wireless and Mobile Computing, Networking and Communications (WiMob), 2015 IEEE 11th International Conference on,* 1–8. IEEE.

Arthur, David, and Sergei Vassilvitskii. 2007. “k-means++: The advantages of careful seeding”. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms,* 1027–1035. Society for Industrial and Applied Mathematics.

Atif, Muhammad, and Serkan Serdaroglu. 2012. “A measurement system for human movement analysis”.

Avci, Akin, Stephan Bosch, Mihai Marin-Perianu, Raluca Marin-Perianu, and Paul Havinga. 2010. “Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey”. In *Architecture of computing systems (ARCS), 2010 23rd international conference on,* 1–10. VDE.

Äyrämö, Sami. 2006. *Knowledge mining using robust clustering.* University of Jyväskylä.

Ball, Geoffrey H, and David J Hall. 1965. *ISODATA, a novel method of data analysis and pattern classification.* Technical report. DTIC Document.

Bamberg, Stacy J Morris, Ari Y Benbasat, Donna Moxley Scarborough, David E Krebs, and Joseph A Paradiso. 2008. “Gait analysis using a shoe-integrated wireless sensor system”. *IEEE transactions on information technology in biomedicine* 12 (4): 413–423.

Bao, Ling, and Stephen S Intille. 2004. “Activity recognition from user-annotated acceleration data”. In *International Conference on Pervasive Computing,* 1–17. Springer.

Bao, Min-Hang. 2000. *Micro mechanical transducers: pressure sensors, accelerometers and gyroscopes.* Volume 8. Elsevier.

Batista, Gustavo EAPA, and Maria Carolina Monard. 2003. “An analysis of four missing data treatment methods for supervised learning”. *Applied artificial intelligence* 17 (5-6): 519–533.

Bayat, Akram, Marc Pomplun, and Duc A Tran. 2014. “A study on human activity recognition using accelerometer data from smartphones”. *Procedia Computer Science* 34:450–457.

Bezdek, James C, and Nikhil R Pal. 1998. “Some new indexes of cluster validity”. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 28 (3): 301–315.

Blair, Steven N, Michael J LaMonte, and Milton Z Nichaman. 2004. “The evolution of physical activity recommendations: how much is enough?” *The American journal of clinical nutrition* 79 (5): 913S–920S.

Bonomi, Alberto G. 2010. “Physical activity recognition using a wearable accelerometer”. In *Sensing Emotions,* 41–51. Springer.

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

Brezmes, Tomas, Juan-Luis Gorricho, and Josep Cotrina. 2009. “Activity recognition from accelerometer data on a mobile phone”. In *Distributed computing, artificial intelligence, bioinformatics, soft computing, and ambient assisted living,* 796–799. Springer.

Bulling, Andreas, Ulf Blanke, and Bernt Schiele. 2014. “A tutorial on human activity recognition using body-worn inertial sensors”. *ACM Computing Surveys (CSUR)* 46 (3): 33.

Butte, Nancy F, Ulf Ekelund, and Klaas R Westerterp. 2012. “Assessing physical activity using wearable monitors: measures of physical activity”. *Med Sci Sports Exerc* 44 (1 Suppl 1): S5–12.

Caliński, Tadeusz, and Jerzy Harabasz. 1974. “A dendrite method for cluster analysis”. *Communications in Statistics-theory and Methods* 3 (1): 1–27.

Carmines, Edward G, and Richard A Zeller. 1979. *Reliability and validity assessment.* Volume 17. Sage publications.

Caspersen, Carl J, Kenneth E Powell, and Gregory M Christenson. 1985. “Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research.” *Public health reports* 100 (2): 126.

Celebi, M Emre, Hassan A Kingravi, and Patricio A Vela. 2013. “A comparative study of efficient initialization methods for the k-means clustering algorithm”. *Expert Systems with Applications* 40 (1): 200–210.

Chastin, SFM, and MH Granat. 2010. “Methods for objective measure, quantification and analysis of sedentary behaviour and inactivity”. *Gait & posture* 31 (1): 82–86.

Chen, Kong Y, and David R Bassett. 2005. “The technology of accelerometry-based activity monitors: current and future”. *Medicine and science in sports and exercise* 37 (11): S490.

Cord, Aurélien, Christophe Ambroise, and Jean-Pierre Cocquerez. 2006. “Feature selection in robust clustering based on Laplace mixture”. *Pattern Recognition Letters* 27 (6): 627–635.

Cortes, Corinna, and Vladimir Vapnik. 1995. “Support-vector networks”. *Machine learning* 20 (3): 273–297.

Crouter, Scott E, James R Churilla, and David R Bassett Jr. 2006. “Estimating energy expenditure using accelerometers”. *European journal of applied physiology* 98 (6): 601–612.

Davies, David L, and Donald W Bouldin. 1979. “A cluster separation measure”. *IEEE transactions on pattern analysis and machine intelligence,* number 2:224–227.

Davis, Roy B. 1988. “Clinical gait analysis”. *IEEE Engineering in Medicine and Biology Magazine* 7 (3): 35–40.

Denzin, Norman. 1970. “Strategies of multiple triangulation”. *The research act in sociology: A theoretical introduction to sociological method* 297:313.

Dixon, John K. 1979. “Pattern recognition with partly missing data”. *IEEE Transactions on Systems, Man, and Cybernetics* 9 (10): 617–621.

Documentation, MATLAB. 2015. *evalclusters. The MathWorks.*

D’Orazio, Tiziana, and Marco Leo. 2010. “A review of vision-based systems for soccer video analysis”. *Pattern recognition* 43 (8): 2911–2926.

Duda, Richard O, Peter E Hart, and David G Stork. 2012. *Pattern classification.* John Wiley & Sons.

Dunn, Joseph C. 1974. “Well-separated clusters and optimal fuzzy partitions”. *Journal of cybernetics* 4 (1): 95–104.

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.

Estivill-Castro, Vladimir. 2002. “Why so many clustering algorithms: a position paper”. *ACM SIGKDD explorations newsletter* 4 (1): 65–75.

Evenson, Kelly R, Diane J Catellier, Karminder Gill, Kristin S Ondrak, and Robert G McMurray. 2008. “Calibration of two objective measures of physical activity for children”. *Journal of sports sciences* 26 (14): 1557–1565.

Evenson, Kelly R, Michelle M Goto, and Robert D Furberg. 2015. “Systematic review of the validity and reliability of consumer-wearable activity trackers”. *International Journal of Behavioral Nutrition and Physical Activity* 12 (1): 159.

Fayyad, Usama M, Gregory Piatetsky-Shapiro, Padhraic Smyth, et al. 1996. “Knowledge discovery and data mining: towards a unifying framework.” In *KDD,* 96:82–88.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996a. “From data mining to knowledge discovery in databases”. *AI magazine* 17 (3): 37.

———. 1996b. “The KDD process for extracting useful knowledge from volumes of data”. *Communications of the ACM* 39 (11): 27–34.

Ferguson, Ty, Alex V Rowlands, Tim Olds, and Carol Maher. 2015. “The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: a cross-sectional study”. *Int J Behav Nutr Phys Act* 12 (1): 42.

Fisher, Ronald A. 1936. “The use of multiple measurements in taxonomic problems”. *Annals of eugenics* 7 (2): 179–188.

Fränti, Pasi, and Olli Virmajoki. 2006. “Iterative shrinking method for clustering problems”. *Pattern Recognition* 39 (5): 761–775.

Freedson, Patty S, Edward Melanson, and John Sirard. 1998. “Calibration of the Computer Science and Applications, Inc. accelerometer.” *Medicine and science in sports and exercise* 30 (5): 777–781.

Freedson, PATTY, David Pober, and Kathleen F Janz. 2005. “Calibration of accelerometer output for children”. *Medicine and science in sports and exercise* 37 (11): S523.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning.* Volume 1. Springer series in statistics Springer, Berlin.

Gasser, Michael. 2014. “Motion Estimation Using Inertial Sensor Technology with Applications to Sporting Exercises”. *The whole is more than the sum of its parts* 29 (3): 15–24.

Goethals, Bart. 2003. “Survey on frequent pattern mining”. *Univ. of Helsinki:*19.

Haapala, HL, MH Hirvensalo, K Laine, Lauri Laakso, H Hakonen, Taru Lintunen, and TH Tammelin. 2014. “Adolescents’ physical activity at recess and actions to promote a physically active school day in four Finnish schools”. *Health education research* 29 (5): 840–852.

Halkidi, Maria, Yannis Batistakis, and Michalis Vazirgiannis. 2002. “Cluster validity methods: part I”. *ACM Sigmod Record* 31 (2): 40–45.

Hallal, Pedro C, Lars Bo Andersen, Fiona C Bull, Regina Guthold, William Haskell, Ulf Ekelund, Lancet Physical Activity Series Working Group, et al. 2012. “Global physical activity levels: surveillance progress, pitfalls, and prospects”. *The lancet* 380 (9838): 247–257.

Hammer, Michael. 1976. “Error detection in data base systems”. In *Proceedings of the June 7-10, 1976, national computer conference and exposition,* 795–801. ACM.

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

Hand, David J, Heikki Mannila, and Padhraic Smyth. 2001. *Principles of data mining.* MIT press.

Hartmann, Antonia, Susanna Luzi, Kurt Murer, Rob A de Bie, and Eling D de Bruin. 2009. “Concurrent validity of a trunk tri-axial accelerometer system for gait analysis in older adults”. *Gait & posture* 29 (3): 444–448.

Health, United States. Department of, and Human Services. 1996. *Physical activity and health: A report of the Surgeon General.* diane Publishing.

Helajärvi, H, K Pahkala, O Raitakari, T Tammelin, J Viikari, and O Heinonen. 2013. “Istu ja pala - Onko istuminen uusi terveysuhka”. *Duodecim* 129 (7): 51–56.

Hendelman, D, K Miller, C Baggett, E Debold, and P Freedson. 2000. “Validity of accelerometry for the assessment of moderate intensity physical activity in the field.” *Medicine and science in sports and exercise* 32 (9 Suppl): S442–9.

Hill, James O, Holly R Wyatt, George W Reed, and John C Peters. 2003. “Obesity and the environment: where do we go from here?” *Science* 299 (5608): 853–855.

Hillman, Charles H, Keita Kamijo, and Mark Scudder. 2011. “A review of chronic and acute physical activity participation on neuroelectric measures of brain health and cognition during childhood”. *Preventive medicine* 52:S21–S28.

Huber, PJ. 1981. *Robust Statistics, 308 pp.*

Husu, Pauliina, Henri Vähäpyä, and Tommi Vasankari. 2016. “Objectively measured sedentary behavior and physical activity of Finnish 7-to 14-year-old children–associations with perceived health status: a cross-sectional study”. *BMC public health* 16 (1): 338.

Huynh, Tâm, and Bernt Schiele. 2005. “Analyzing features for activity recognition”. In *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies,* 159–163. ACM.

Incel, Ozlem Durmaz, Mustafa Kose, and Cem Ersoy. 2013. “A review and taxonomy of activity recognition on mobile phones”. *BioNanoScience* 3 (2): 145–171.

Intille, Stephen S, Jonathan Lester, James F Sallis, and Glen Duncan. (2012). “New horizons in sensor development”. *Medicine and science in sports and exercise* 44 (1 Suppl 1): S24.

Jain, Anil K. 2010. “Data clustering: 50 years beyond K-means”. *Pattern recognition letters* 31 (8): 651–666.

Jain, Anil K, and Richard C Dubes. 1988. *Algorithms for clustering data.* Prentice-Hall, Inc.

Jain, Anil K, M Narasimha Murty, and Patrick J Flynn. 1999. “Data clustering: a review”. *ACM computing surveys (CSUR)* 31 (3): 264–323.

Jankovic, Joseph. 2008. “Parkinson’s disease: clinical features and diagnosis”. *Journal of Neurology, Neurosurgery & Psychiatry* 79 (4): 368–376.

Jans, Marielle P, Karin I Proper, and Vincent H Hildebrandt. 2007. “Sedentary behavior in Dutch workers: differences between occupations and business sectors”. *American journal of preventive medicine* 33 (6): 450–454.

Jauhiainen, Susanne, and Tommi Kärkkäinen. 2017. “A Simple Cluster Validation Index with Maximal Coverage”. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESAINN 2017,* 293–298.

Jiang, Shanshan, Yanchuan Cao, Sameer Iyengar, Philip Kuryloski, Roozbeh Jafari, Yuan Xue, Ruzena Bajcsy, and Stephen Wicker. 2008. "CareNet: an integrated wireless sensor networking environment for remote healthcare". In *Proceedings of the ICST 3rd international conference on Body area networks,* 9. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Karantonis, Dean M, Michael R Narayanan, Merryn Mathie, Nigel H Lovell, and Branko G Celler. 2006. "Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring". *IEEE transactions on information technology in biomedicine* 10 (1): 156–167.

Kärkkäinen, Tommi. 2015. "Assessment of Feature Saliency of MLP using Analytic Sensitivity". In *Proceedings,* 273. Presses universitaires de Louvain.

Kärkkäinen, Tommi, and Erkki Heikkola. 2004. "Robust formulations for training multilayer perceptrons". *Neural Computation* 16 (4): 837–862.

Kärkkäinen, Tommi, and Mirka Saarela. 2015. "Robust principal component analysis of data with missing values". In *International Workshop on Machine Learning and Data Mining in Pattern Recognition,* 140–154. Springer.

Khan, Zafar A, and Won Sohn. 2011. "Abnormal human activity recognition system based on R-transform and kernel discriminant technique for elderly home care". *IEEE Transactions on Consumer Electronics* 57 (4).

Kim, Eunju, Sumi Helal, and Diane Cook. 2010. "Human activity recognition and pattern discovery". *IEEE Pervasive Computing* 9 (1): 48–53.

Kim, Minho, and RS Ramakrishna. 2005. "New indices for cluster validity assessment". *Pattern Recognition Letters* 26 (15): 2353–2363.

Kim, Youngdeok, Gregory J Welk, Saori I Braun, and Minsoo Kang. 2015. "Extracting objective estimates of sedentary behavior from accelerometer data: measurement considerations for surveillance and research applications". *PloS one* 10 (2): e0118078.

Kriegel, Hans-Peter, Peer Kröger, Jörg Sander, and Arthur Zimek. 2011. "Density-based clustering". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1 (3): 231–240.

Kwapisz, Jennifer R, Gary M Weiss, and Samuel A Moore. 2011. "Activity recognition using cell phone accelerometers". *ACM SigKDD Explorations Newsletter* 12 (2): 74–82.

Lara, Oscar D, and Miguel A Labrador. 2013. "A survey on human activity recognition using wearable sensors". *IEEE Communications Surveys & Tutorials* 15 (3): 1192–1209.

Lee, Jung-Min, Youngwon Kim, and Gregory J Welk. 2014. "Validity of consumer-based physical activity monitors". *Med Sci Sports Exerc* 46 (9): 1840–8.

Little, Roderick J, and Donald B Rubin. 1987. *Statistical analysis with missing data.* John Wiley & Sons.

Little, Roderick JA, and Donald B Rubin. 1989. "The analysis of social science data with missing values". *Sociological Methods & Research* 18 (2-3): 292–326.

———. 2014. *Statistical analysis with missing data.* John Wiley & Sons.

Liu, Huan, and Hiroshi Motoda. 1998. *Feature extraction, construction and selection: A data mining perspective.* Volume 453. Springer Science & Business Media.

Liu, Yanchi, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. 2010. "Understanding of internal clustering validation measures". In *Data Mining (ICDM), 2010 IEEE 10th International Conference on,* 911–916. IEEE.

Lotfi, Ahmad, Minh Nguyen, and Caroline Langensiepen. 2015. "Human gait classification using a tri-axial accelerometer". In *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments,* 36. ACM.

Luo, Suhuai, and Qingmao Hu. 2004. "A dynamic motion pattern analysis approach to fall detection". In *Biomedical Circuits and Systems, 2004 IEEE International Workshop on,* 1–5. IEEE.

Mabroukeh, Nizar R, and Christie I Ezeife. 2010. "A taxonomy of sequential pattern mining algorithms". *ACM Computing Surveys (CSUR)* 43 (1): 3.

MacQueen, James, et al. 1967. “Some methods for classification and analysis of multivariate observations”. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability,* 1:281–297. 14. Oakland, CA, USA.

Mannini, Andrea, Stephen S Intille, Mary Rosenberger, Angelo M Sabatini, and William Haskell. 2013. “Activity recognition using a single accelerometer placed at the wrist or ankle”. *Medicine and science in sports and exercise* 45 (11): 2193.

Marquis, Scott, M Milar Moore, Diane B Howieson, Gary Sexton, Haydeh Payami, Jeffrey A Kaye, and Richard Camicioli. 2002. “Independent predictors of cognitive decline in healthy elderly persons”. *Archives of neurology* 59 (4): 601–606.

Montgomery, Douglas C, Elizabeth A Peck, and G Geoffrey Vining. 2015. *Introduction to linear regression analysis.* John Wiley & Sons.

Morales Gonzalez, Juan Francisco. 2015. “Development of a Real-time System to Measure Foot Clearance Based on Inertial Sensors”. Masters thesis, Institute of Biomaterials and Biomedical Engineering, University of Toronto.

Morin, JB, P Samozino, K Zameziati, and A Belli. 2007. “Effects of altered stride frequency and contact time on leg-spring behavior in human running”. *Journal of biomechanics* 40 (15): 3341–3348.

Mubashir, Muhammad, Ling Shao, and Luke Seed. 2013. “A survey on fall detection: Principles and approaches”. *Neurocomputing* 100:144–152.

Najafi, Bijan, Kamiar Aminian, Anisoara Paraschiv-Ionescu, François Loew, Christophe J Bula, and Philippe Robert. 2003. “Ambulatory system for human motion analysis using a kinematic sensor: monitoring of daily physical activity in the elderly”. *IEEE Transactions on biomedical Engineering* 50 (6): 711–723.

Nelson, M Benjamin, Leonard A Kaminsky, D Clark Dickin, and AH Montoye. (2016). “Validity of Consumer-Based Physical Activity Monitors for Specific Activity Types.” *Medicine and science in sports and exercise.*

Nieminen, Paavo, Ilkka Pölönen, and Tuomo Sipola. 2013. “Research literature clustering using diffusion maps”. *Journal of Informetrics* 7 (4): 874–886.

Opetusministeriö & Nuori Suomi. 2008. “Fyysisen aktiivisuuden suositus kouluikäisille 7–18-vuotiaille”.

Organization, World Health. 2010. *Global recommendations on Physical Activity for health.* World Health Organization.

Owen, Neville, Phillip B Sparling, Geneviève N Healy, David W Dunstan, and Charles E Matthews. 2010. “Sedentary behavior: emerging evidence for a new health risk”. In *Mayo Clinic Proceedings,* 85:1138–1141. 12. Elsevier.

Pakhira, Malay K, Sanghamitra Bandyopadhyay, and Ujjwal Maulik. 2004. “Validity index for crisp and fuzzy clusters”. *Pattern recognition* 37 (3): 487–501.

Pate, Russell R, Jonathan A Mitchell, Wonwoo Byun, and Marsha Dowda. 2011. “Sedentary behaviour in youth”. *British journal of sports medicine* 45 (11): 906–913.

Pearson, Karl. 1901. “LIII. On lines and planes of closest fit to systems of points in space”. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2 (11): 559–572.

Pentland, Alex. 2000. “Looking at people: Sensing for ubiquitous and wearable computing”. *IEEE Transactions on Pattern analysis and machine intelligence* 22 (1): 107–119.

Petrosyan, Vahan, and Alexandre Proutiere. 2016. “Viral Clustering: a robust method to extract structures in heterogeneous datasets”. In *The Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16).*

Piatetsky-Shapiro, Gregory. 1990. “Knowledge discovery in real databases: A report on the IJCAI-89 Workshop”. *AI magazine* 11 (4): 68.

———. 2000. “Knowledge discovery in databases: 10 years after”. *ACM SIGKDD Explorations Newsletter* 1 (2): 59–61.

Poikola, KK Antti, and H Honko. 2010. *Mydata a nordic model for human-centered personal data management and processing.* Technical report. tech. rep., Ministry of Transport Finland.

Poppe, Ronald. 2007. “Vision-based human motion analysis: An overview”. *Computer vision and image understanding* 108 (1): 4–18.

Preece, Stephen J, John Y Goulermas, Laurence PJ Kenney, Dave Howard, Kenneth Meijer, and Robin Crompton. 2009. “Activity identification using body-mounted sensors—a review of classification techniques”. *Physiological measurement* 30 (4): R1.

Ray, Siddheswar, and Rose H Turi. 1999. “Determination of number of clusters in k-means clustering and application in colour image segmentation”. In *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques,* 137–143. Calcutta, India.

Rendón, Eréndira, Itzel Abundez, Alejandra Arizmendi, and ElviaM Quiroz. 2011. “Internal versus external cluster validation indexes”. *International Journal of computers and communications* 5 (1): 27–34.

Rezende, Leandro Fornias Machado de, Maurício Rodrigues Lopes, Juan Pablo Rey-López, Victor Keihan Rodrigues Matsudo, and Olinda do Carmo Luiz. 2014. “Sedentary behavior and health outcomes: an overview of systematic reviews”. *PloS one* 9 (8).

Rokach, Lior, and Oded Maimon. 2014. *Data mining with decision trees: theory and applications.* World scientific.

Rousseeuw, Peter J. 1987. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. *Journal of computational and applied mathematics* 20:53–65.

Rueterbories, Jan, Erika G Spaich, Birgit Larsen, and Ole K Andersen. 2010. “Methods for gait event detection and analysis in ambulatory systems”. *Medical engineering & physics* 32 (6): 545–552.

Saarela, Mirka, Joonas Hämäläinen, and Tommi Kärkkäinen. 2017. “Feature Ranking of Large, Robust, and Weighted Clustering Result”. In *Proceedings of 21st Pacific Asia Conference on Knowledge Discovery and Data Mining - PAKDD 2017.* To appear (12 pages).

Saarela, Mirka, and Tommi Kärkkäinen. 2015. “Analysing student performance using sparse data of core bachelor courses”. *JEDM-Journal of Educational Data Mining* 7 (1): 3–32.

Salmon, JO, Kylie Ball, David Crawford, Michael Booth, Amanda Telford, Clare Hume, Damien Jolley, and Anthony Worsley. 2005. "Reducing sedentary behaviour and increasing physical activity among 10-year-old children: overview and process evaluation of the 'Switch-Play'intervention". *Health promotion international* 20 (1): 7–17.

Sasaki, Jeffer E, Dinesh John, and Patty S Freedson. 2011. "Validation and comparison of ActiGraph activity monitors". *Journal of Science and Medicine in Sport* 14 (5): 411–416.

Schoeller, DA, and E Van Santen. 1982. "Measurement of energy expenditure in humans by doubly labeled water method". *Journal of Applied Physiology* 53 (4): 955–959.

Simon, Sheldon R. 2004. "Quantification of human motion: gait analysis—benefits and limitations to its application to clinical problems". *Journal of biomechanics* 37 (12): 1869–1880.

Sprent, Peter, and Nigel C Smeeton. 2016. *Applied nonparametric statistical methods.* CRC Press.

Stone, Erik E, and Marjorie Skubic. 2011. "Passive in-home measurement of stride-to-stride gait variability comparing vision and Kinect sensing". In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE,* 6491–6494. IEEE.

Swan, Melanie. 2013. "The quantified self: Fundamental disruption in big data science and biological discovery". *Big Data* 1 (2): 85–99.

Syväoja, Heidi, Marko T Kantomaa, Timo Ahonen, Harto Hakonen, Anna Kankaanpää, and Tuija H Tammelin. 2013. "Physical activity, sedentary behavior, and academic performance in Finnish children".

Tammelin, Tuija, Ulf Ekelund, Jouko Remes, and Simo Näyhä. 2007. "Physical activity and sedentary behaviors among Finnish youth." *Medicine and science in sports and exercise* 39 (7): 1067–1074.

Tammelin, Tuija, Kaarlo Laine, and Salla Turpeinen. 2012. "Liikkuva koulu-ohjelman pilottivaiheen 2010–2012 loppuraportti". *Liikunnan ja kansanterveyden julkaisuja* 261.

Tan, Vipin, Pang-Ning Kumar, and Michael Steinbach. 2006. *Introduction to data mining.* Pearson Education India.

Tanasa, Doru, and Brigitte Trousse. 2004. “Advanced data preprocessing for intersites web usage mining”. *IEEE Intelligent Systems* 19 (2): 59–65.

Tapia, Emmanuel Munguia, Stephen S Intille, and Kent Larson. 2004. “Activity recognition in the home using simple and ubiquitous sensors”. In *International Conference on Pervasive Computing,* 158–175. Springer.

Thorndike, Robert L. 1953. “Who belongs in the family?” *Psychometrika* 18 (4): 267–276.

Tong, Kaiyu, and Malcolm H Granat. 1999. “A practical gait analysis system using gyroscopes”. *Medical engineering & physics* 21 (2): 87–94.

Trost, Stewart G, Paul D Loprinzi, Rebecca Moore, and Karin A Pfeiffer. 2011. “Comparison of accelerometer cut points for predicting activity intensity in youth”. *Med Sci Sports Exerc* 43 (7): 1360–1368.

Tukey, John W. 1980. “We need both exploratory and confirmatory”. *The American Statistician* 34 (1): 23–25.

Ustev, Yunus Emre, Ozlem Durmaz Incel, and Cem Ersoy. 2013. “User, device and orientation independent human activity recognition on mobile phones: Challenges and a proposal”. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication,* 1427–1436. ACM.

Varhaisvuosien fyysisen aktiivisuuden suositukset. 2016. “Iloa, leikkiä ja yhdessä tekemistä.” *Opetus- ja kulttuuriministeriö 2016:21.*

Verleysen, Michel, and Damien François. 2005. “The curse of dimensionality in data mining and time series prediction”. In *International Work-Conference on Artificial Neural Networks,* 758–770. Springer.

Vesterinen, V, K Häkkinen, E Hynynen, J Mikkola, L Hokka, and A Nummela. 2013. “Heart rate variability in prediction of individual adaptation to endurance training in recreational endurance runners”. *Scandinavian journal of medicine & science in sports* 23 (2): 171–180.

Vesterinen, Ville. 2016. *Predicting and Monitoring Individual Endurance Training Adaptation and Individualizing Training Prescription.* University of Jyväskylä.

Wartiainen, Pekka, and Tommi Kärkkäinen. 2015. “Hierarchical, prototype-based clustering of multiple time series with missing values”. In *ESANN 2015: Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning,* 95–100.

Weyand, Peter G, Deborah B Sternlight, Matthew J Bellizzi, and Seth Wright. 2000. “Faster top running speeds are achieved with greater ground forces not more rapid leg movements”. *Journal of applied physiology* 89 (5): 1991–1999.

Whittle, Michael W. 2014. *Gait analysis: an introduction.* Butterworth-Heinemann.

WHO. 2009. *Global health risks: mortality and burden of disease attributable to selected major risks.* World Health Organization.

Xanthopoulos, Petros, Panos M Pardalos, and Theodore B Trafalis. 2013. “Linear discriminant analysis”. In *Robust Data Mining,* 27–33. Springer.

Xu, Rui, and Donald Wunsch. 2005. “Survey of clustering algorithms”. *IEEE Transactions on neural networks* 16 (3): 645–678.

Yang, Che-Chang, and Yeh-Liang Hsu. 2010. “A review of accelerometry-based wearable motion detectors for physical activity monitoring”. *Sensors* 10 (8): 7772–7788.

Yang, Jhun-Ying, Jeen-Shing Wang, and Yen-Ping Chen. 2008. “Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers”. *Pattern recognition letters* 29 (16): 2213–2220.

Zaki, Mohammed J, and Wagner Meira Jr. 2014. *Data mining and analysis: fundamental concepts and algorithms.* Cambridge University Press.