OULU 2015
C 542

ACTA UNIVERSITATIS OULUENSIS

Eija Ferreira

MODEL SELECTION IN TIME SERIES MACHINE LEARNING APPLICATIONS

*UNIVERSITY OF OULU GRADUATE SCHOOL;
UNIVERSITY OF OULU,
FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING;
INFOTECH OULU*

ACTA UNIVERSITATIS OULUENSIS
C Technica 542

EIJA FERREIRA

MODEL SELECTION IN TIME SERIES MACHINE LEARNING APPLICATIONS

Academic dissertation to be presented with the assent of the Doctoral Training Committee of Technology and Natural Sciences of the University of Oulu for public defence in Auditorium TS101, Linnanmaa, on 11 September 2015, at 12 noon

UNIVERSITY OF OULU, OULU 2015

Acta Univ. Oul. C 542, 2015

Supervised by
Professor Juha Röning
Professor Anind K. Dey

Reviewed by
Professor Xiaohui Liu
Professor Barbara Hammer

Opponent
Associate Professor Daniel Roggen

ISBN 978-952-62-0900-5 (Paperback)
ISBN 978-952-62-0901-2 (PDF)

ISSN 0355-3213 (Printed)
ISSN 1796-2226 (Online)

Cover Design
Raimo Ahonen

JUVENES PRINT
TAMPERE 2015

Ferreira, Eija, Model selection in time series machine learning applications.
University of Oulu Graduate School; University of Oulu, Faculty of Information Technology and Electrical Engineering, Department of Computer Science and Engineering; Infotech Oulu
Acta Univ. Oul. C 542, 2015
University of Oulu, P.O. Box 8000, FI-90014 University of Oulu, Finland

Abstract

Model selection is a necessary step for any practical modeling task. Since the true model behind a real-world process cannot be known, the goal of model selection is to find the best approximation among a set of candidate models.

In this thesis, we discuss model selection in the context of time series machine learning applications. We cover four steps of the commonly followed machine learning process: data preparation, algorithm choice, feature selection and validation. We consider how the characteristics and the amount of data available should guide the selection of algorithms to be used, and how the data set at hand should be divided for model training, selection and validation to optimize the generalizability and future performance of the model. We also consider what are the special restrictions and requirements that need to be taken into account when applying regular machine learning algorithms to time series data. We especially aim to bring forth problems relating model over-fitting and over-selection that might occur due to careless or uninformed application of model selection methods.

We present our results in three different time series machine learning application areas: resistance spot welding, exercise energy expenditure estimation and cognitive load modeling. Based on our findings in these studies, we draw general guidelines on which points to consider when starting to solve a new machine learning problem from the point of view of data characteristics, amount of data, computational resources and possible time series nature of the problem. We also discuss how the practical aspects and requirements set by the environment where the final model will be implemented affect the choice of algorithms to use.

Keywords: machine learning, model selection, real-world applications, time series data

Ferreira, Eija, Aikasarjatiedon mallinnus ja mallinvalinta koneoppimissovelluksissa.
Oulun yliopiston tutkijakoulu; Oulun yliopisto, Tieto- ja sähkötekniikan tiedekunta, Tietotekniikan osasto; Infotech Oulu

Acta Univ. Oul. C 542, 2015
Oulun yliopisto, PL 8000, 90014 Oulun yliopisto

Tiivistelmä

Mallinvalinta on oleellinen osa minkä tahansa käytännön mallinnusongelman ratkaisua. Koska mallinnettavan ilmiön toiminnan taustalla olevaa todellista mallia ei voida tietää, on mallinvalinnan tarkoituksena valita malliehdokkaiden joukosta sitä lähimpänä oleva malli.

Tässä väitöskirjassa käsitellään mallinvalintaa aikasarjamuotoista dataa sisältävissä sovelluksissa neljän koneoppimisprosessissa yleisesti noudatetun askeleen kautta: aineiston esikäsittely, algoritmin valinta, piirteiden valinta ja validointi. Väitöskirjassa tutkitaan, kuinka käytettävissä olevan aineiston ominaisuudet ja määrä tulisi ottaa huomioon algoritmin valinnassa, ja kuinka aineisto tulisi jakaa mallin opetusta, testausta ja validointia varten mallin yleistettävyyden ja tulevan suorituskyvyn optimoimiseksi. Myös erityisiä rajoitteita ja vaatimuksia tavanomaisten koneoppimismenetelmien soveltamiselle aikasarjadataan käsitellään. Työn tavoitteena on erityisesti tuoda esille mallin ylioppimiseen ja ylivalintaan liittyviä ongelmia, jotka voivat seurata mallinvalin- tamenetelmien huolimattomasta tai osaamattomasta käytöstä.

Työn käytännön tulokset perustuvat koneoppimismenetelmien soveltamiseen aikasar- jadatan mallinnukseen kolmella eri tutkimusalueella: pistehitsaus, fyysisen harjoittelun aikasen energiankulutuksen arviointi sekä kognitiivisen kuormituksen mallintaminen. Väitöskirja tarjoaa näihin tuloksiin pohjautuen yleisiä suuntaviivoja, joita voidaan käyttää apuna lähdettäessä ratkaisemaan uutta koneoppimisongelmaa erityisesti aineiston ominaisuuksien ja määrän, laskennallisten resurssien sekä ongelman mahdollisen aikasar- jaluonteen näkökulmasta. Työssä pohditaan myös mallin lopullisen toimintaympäristön asettamien käytännön näkökohtien ja rajoitteiden vaikutusta algoritmin valintaan.

Asiasanat: aikasarjadata, koneoppiminen, käytännön sovellukset, mallinvalinta

To my Denzil and all of my family

Acknowledgements

I would like to start by thanking my two supervisors Professor Juha Röning (University of Oulu, Finland) and Professor Anind K. Dey (Carnegie Mellon University, USA) for giving me the opportunity to work toward a Ph.D. in their research groups, and for their support along the way. I would especially like to thank Professor Röning for his trust and appreciation for my work, and also for the freedom he has given me to pursue my research. I would like to thank Professor Dey for welcoming me into his research group, and for treating me like family. Thank you for the countless hours you dedicated to my research, for all the encouragement, and appreciation of my ideas.

I am grateful to the reviewers of the thesis manuscript, Professor Xiaohui Liu (Brunel University, UK) and Professor Barbara Hammer (Bielefeld University, Germany), for their thoughtful and valuable comments. I would like to thank Reader Daniel Roggen (University of Sussex, UK) for accepting the opponent role at my thesis defence.

I would like to acknowledge all my colleagues and co-authors of the related articles. I thank the members and wannabe members of the Data Analysis and Inference research group at the University of Oulu for the positive, supportive and fun work environment. I feel really privileged to be able to work with my close friends. I especially want to thank Dr. Heli Koskimäki for all the good advice on research and non-research related topics, and Dr. Pekka Siirtola for his patience with me frequently popping into his office with random questions. I also thank the members of the Ubicomp Lab at Carnegie Mellon University for introducing me to a whole new way of doing research, and for sharing many moments in and out of the office.

This work was financially supported by several organizations. I gratefully acknowledge the support from the Graduate School in Electronics, Telecommunications and Automation (GETA), Tekniikan edistämissäätiö, Tauno Tönning Foundation, Nokia Foundation, Emil Aaltonen Foundation, Seppo Säynäjäkangas Science Foundation, Oulu University Scholarship Foundation, and the University of Oulu Graduate School. I would especially like to express my gratitude to the Binational Fulbright Commission in Helsinki, Finland, and the Technology Industries of Finland Centennial Foundation for awarding me the Fulbright-Technology Industries of Finland Grant for a research visit in the United States. This experience completely changed my view of the academic world, and the world in general. The projects in which I have created contributions for

this thesis received funding from the European Commission and the Finnish Funding Agency for Innovation. I would like to thank all the industrial partners and the people I have collaborated with in the projects.

Finally, I would like to thank my friends and family for all their support over the years. Leena ja Erkki, kiitos taustajoukoille kaikesta tuesta ja kannustuksesta! Most importantly, I thank my dear husband, Denzil, for always being there for me, always believing in me, and always finding new ways to make my day, and our son Toivo for making me appreciate every fleeting moment of my life.

Oulu, August 2015 Eija Ferreira

Symbols and abbreviations

Mathematical notations

β	*p-dimensional vector of regression coefficients*
c_j	*Class j*
ε	*N-dimensional vector of random errors*
$g(x)$	*Classification rule*
k	*Number of nearest neighbors*
λ_j	*j*th *eigenvalue*
m	*Number of classes*
N	*Sample size*
n	*Number of features*
p	*Dimension of feature space*
$P(c_j)$	*A priori probability of class* c_j
$P(c_j \mid x)$	*A posteriori probability of class* c_j
$p(x)$	*Probability density function*
$p(x \mid c_j)$	*Class-conditional probability density function*
u_j	*j*th *eigenvector*
x	*Input vector*
X	*($N \times p$)-dimensional matrix of explanatory variables*
x_i	*Input variable*
y	*Response variable*
Y	*N-dimensional vector of responses*

Abbreviations

BR	*Breathing rate*
ECG	*Electrocardiography*
ECT	*Elementary cognitive task*
EE	*Energy expenditure*
EEG	*Electroencephalography*
EMG	*Electromyography*
EOG	*Electrooculography*

FA	*Finding A’s test*
GC	*Gestalt completion test*
GLVQ	*Generalized learning vector quantization*
GSR	*Galvanic skin response*
HP	*Hidden pattern test*
HR	*Heart rate*
HRV	*Heart rate variability*
HWH	*Harms+Wende GmbH & Co.KG*
IQR	*Interquartile range*
*k*NN	*k-nearest neighbors classifier*
LDA	*Linear discriminant analysis*
LVQ	*Learning vector quantization*
MAD	*Median absolute deviation*
MAFD	*Mean of absolute values of first differences*
MASD	*Mean of absolute values of second differences*
MD	*Mahalanobis discrimination*
MET	*Metabolic equivalent*
MSE	*Mean squared error*
*n*Best	*n best features selection*
NB	*Naïve Bayes classifier*
NC	*Number comparison test*
OLVQ	*Optimized learning rate learning vector quantization*
PCA	*Principal component analysis*
pNN20	*Relative occurrence of successive R-R differences exceeding 20ms*
pNN50	*Relative occurrence of successive R-R differences exceeding 50ms*
PT	*Pursuit test*
QDA	*Quadratic discriminant analysis*
R-R	*Heartbeat R wave to R wave interval*
RMSSD	*Root mean square of successive differences of R-R intervals*
RSW	*Resistance spot welding*
SBFS	*Sequential backward floating selection*
SBS	*Sequential backward selection*
SBT	*Stanzbiegetechnik GES.M.B.H.*
SCR	*Skin conductance response*
SDNN	*Standard deviation of R-R intervals*

SFFS	*Sequential forward floating selection*
SFS	*Sequential forward selection*
SLVQ	*Soft learning vector quantization*
SOM	*Self-organizing map*
SX	*Scattered X's test*

List of original publications

This thesis is based on the following publications, which are referred to in the text by their Roman numerals (I–V):

I Haapalainen E*, Laurinen P, Junno H, Tuovinen L & Röning J (2005) Methods for classifying spot welding processes: a comparative study of performance. Proc. 18th International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems, 412-421.

II Haapalainen E*, Laurinen P, Junno H, Tuovinen L & Röning J (2008) Feature selection for identification of spot welding processes. In: Andrade-Cetto J, Ferrier J-L, Pereira JD, Filipe J (eds) Informatics in Control Automation and Robotics, volume 15 of *Lecture Notes in Electrical Engineering*, 69-79. Springer, Heidelberg.

III Haapalainen E*, Laurinen P, Siirtola P, Röning J, Kinnunen H & Jurvelin H (2008) Exercise energy expenditure estimation based on acceleration data using the linear mixed model. Proc. IEEE International Conference on Information Reuse and Integration, 131-136.

IV Haapalainen E*, Kim S, Forlizzi J & Dey A (2010) Psycho-physiological measures for assessing cognitive load. Proc. 12th ACM International Conference on Ubiquitous Computing, 301-310.

V Ferreira E, Ferreira D, Kim S, Siirtola P, Röning J, Forlizzi J & Dey A (2014) Assessing real-time cognitive load based on psycho-physiological measures for younger and older adults. Proc. IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain, 39-48.

*Eija Ferreira (née Haapalainen)

Contents

1 Introduction

Essentially, all models are wrong, but some are useful.

– George E. P. Box, 1987

These famous words from statistician George E. P. Box summarize the essence of model selection. Since it is in practice impossible to know the true model behind any real-world process, the goal of model selection is to find the one that provides the most useful approximation.

The goal of *modeling* is to derive new information based on the data available, to describe the data by summarizing it in some appropriate way, or to make predictions about future data values (Hand *et al.* 2001). Models are used across the fields of science ranging from computer science to physics, and economics to engineering. The principal idea of *model selection* is to estimate the performance of different model candidates in order to choose the best model achievable (Hastie *et al.* 2009). The improved model performance achieved with model selection often brings along more reliable functioning of the model, better predictions about future outcomes of the process, financial savings or increased safety in safety-critical applications.

In this thesis, we will discuss model selection in the context of time series data. A *time series* is a sequence of observations that are measured at specified times, typically at uniform intervals. Even though the phenomena being measured usually are continuous by nature, the measurements need to be made at discrete time steps for representation in computer memory. The measurements at each time can be multivariate. Because of the unique structure, and especially the high correlation between the subsequent observations of a time series, caution needs to be taken when applying traditional modeling algorithms to time series data.

We will discuss model selection in the field of machine learning. There are, however, three interlinked research areas where our results equally apply: *machine learning*, *pattern recognition* and *data mining*. Machine learning is concerned with the question of how to construct computer algorithms that automatically improve through experience

(Mitchell 1997). Bishop (2006) defines pattern recognition as follows: "The field of pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories." It is difficult to tell the difference between pattern recognition and machine learning since the fields are closely related and use similar methods. This is explained by the fact that these activities are in fact two facets of a field that is the product of two different disciplines. Pattern recognition has its origins in engineering, while machine learning grew out of computer science (Bishop 2006).

Data mining is also closely related to the aforementioned fields. Hand *et al.* (2001) define it as follows: "Data mining is 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." Jain *et al.* (2000) point out that even though it is often difficult to teach a machine to solve a pattern recognition problem, the solution may be trivial to the human eye. In data mining, however, the pattern is often impossible for a human to perceive, since it is identified in millions of patterns. In this case, the computational implementation of the problem is also more demanding. In the following, we will consistently use the term machine learning even if the methods or considerations discussed were originally presented in the context of any of the other two disciplines.

1.1 Objectives and scope

This thesis aims to discuss aspects of the model selection process that should be considered when solving a machine learning problem in general, and when working with time series data in particular. We cover questions including selection of the model type and the feature selection strategy, taking into consideration the characteristics of the dataset at hand. We discuss the division of the dataset for training and testing the models, with additional considerations for modeling temporally dependent time series data. We also consider how the temporal dependency of time series data affects the selection of model type. From our experience in model selection within three different machine learning application areas, all involving time series data, we draw general guidelines, also supported by literature, which can be taken as a starting point when tackling a new machine learning problem. We emphasize the special caution that needs to be addressed when working with time series data.

In summary, we will consider the following research questions in this thesis:

1. How should the data characteristics and the amount of data available guide the selection of model type and the model selection strategy?
2. How should the data set available be used for model training, selection and validation to optimize the model generalizability and performance on new data?
3. What are the special considerations to take into account when applying machine learning algorithms to time series data?

1.2 Contributions of the thesis

The scientific contributions of the individual publications in this compilation lie in their individual application areas: in how these real-world modeling problems are solved using machine learning methods. The contribution of this thesis, however, comprises of an understanding of how to select the methods based on the nature of the problem at hand and the data set available. While these insights are not novel individually, combined they give guidance and help in avoiding errors that are commonly made during the model selection process (Reunanen 2003, Grimes *et al.* 2008).

Publications I and II present a method to classify resistance spot welding (RSW) processes to be able to use information from previous welding batches when setting up a new batch with similar characteristics. In Publication I, we searched for suitable features calculated from welding signals to present their characteristics and looked for a classifier that would give the best results in classifying welding processes. We compared both parametric and nonparametric methods to classify this data set. The thesis author's contributions in this publication included data analysis and interpretation of the results.

In Publication II, we further explored the feature space to eliminate features with less classification-relevant information to speed up the classification and to improve the classification accuracy. We compared a number of search strategies for the feature selection problem. In this publication, the author's contributions included planning and implementing the data analysis as well as interpreting the results.

Publication III presets our results on human exercise energy expenditure (EE) estimation, based on measurements from a wrist-worn acceleration sensor. We introduced a method of modeling exercise EE using the linear mixed model that takes into account the temporal dependencies in the data. The author's contributions included selecting the model type to be used, implementing the models and interpreting the results, while the study was designed together with the other authors.

The last two publications (Publications IV and V) discuss modeling the human cognitive load based on psycho-physiological measurements. In Publication IV, we collected data from a range of psycho-physiological sensors from young adults and extracted a large number of features from the measured signals to explore which ones are useful for assessing cognitive load. The thesis author's contributions consisted of pre-processing the data, implementing the models and analyzing the results, while the study was designed and the data collected mainly by the other authors.

In Publication V, we further explored if we could build cognitive load models also for older adults, and if it would be the same psycho-physiological measurements that are indicative of cognitive load for older adults as for young adults. The thesis author was mainly responsible for planning and running the study with her contributions including data collection and pre-processing, modeling and interpretation of the results.

1.3 Outline of the thesis

The rest of this thesis is organized as follows. In Chapter 2, we introduce the theoretical background and review the related literature for our work. We start by defining the machine learning process commonly followed when solving a modeling problem, and then focus on four individual steps of the process: data preparation, selection of the model type, feature selection and validation. We finish the chapter with a discussion of the special requirements set by the nature of time series data. In Chapter 3, we introduce the results of the contributing publications and summarize their contributions to this thesis. In Chapter 4, we discuss the contributions of our work as a whole from both a theoretical and a practical point of view. Chapter 5 wraps up the thesis with final conclusions.

2 Theoretical foundation

This chapter introduces the theoretical background for our work, as well as the methods we used. Section 2.1 describes the modeling process commonly used when solving a machine learning problem. Section 2.2 presents the data preparation methods, and Section 2.3 the models we used. Section 2.4 focuses on algorithms and procedures we applied for model selection. All of these methods are commonly used in machine learning applications, but they are also suitable for modeling time series data as long as they are applied with some caution. These considerations are brought forth in Section 2.5. We will discuss the choice of methods for this work and their implications for our results later in Chapter 4.

2.1 Modeling process

The sequence of steps commonly followed when solving a machine learning problem is known as the *machine learning process* (Marsland 2014) or equivalently the *data mining process* (Kantardzic 2011). It is "the process by which machine learning algorithms can be selected, applied, and evaluated for the problem" (Marsland 2014). The terms used in the literature for the different steps vary somewhat, while the tasks of the process remain mostly the same. Marsland (2014) defines the steps of the highly iterative process as follows:

1. Data collection and preparation
2. Feature selection
3. Algorithm choice
4. Parameter and model selection
5. Training
6. Evaluation

Hand *et al.* (2001) mention three similar steps: selection of the model structure, the criterion function and the optimization strategy. They emphasize the importance of tailoring each of the algorithms used for the specific application at hand rather than using a commonly used combination.

In this thesis, we will concentrate especially on the tasks of data preparation, selection of the model type, feature selection and validation. We use the term *model* *selection* with a broader meaning to cover all the aspects of model structure, parameter and feature selection.

2.2 Data preparation

According to (Pyle 1999), the purpose of data preparation is to obtain a data set that produces better models faster than could be done with unprepared data. This should be done in a way that does not disturb the natural order of the data but best enhances the data for the particular task at hand. In this work, we have applied *outlier detection*, *feature extraction* as well as two types of data preparation techniques that rescale the range of a variable: data *normalization* and data *standardization*.

2.2.1 Outlier detection and missing values

An *outlier* is defined as a very low frequency occurrence of a value of a variable that is located far away from the mean of the variable (Pyle 1999). Outliers are often caused by measurement error due to device malfunction or an error in reading the data. If the plausible range of measurement values is known, values falling outside this range can be pointed out. If, however, the true range of the values is not known, a threshold for an acceptable distance from the mean can be used, usually proportional to the standard deviation of the distribution of the variable. For a normally distributed random variable, 95% of the data points lie within two standard deviations from the mean, and 99% of the values are located within three standard deviations (Theodoridis & Koutroumbas 2008).

Most modeling methods are sensitive to outliers in data. Two approaches for handling outliers are commonly used. If the number of outliers is very small, they are often discarded. However, many modeling methods are unable to utilize observations where some of the variable values are missing. Therefore, the whole multi-dimensional observation needs to be discarded. This reduces the amount of information available for modeling and can have a negative influence on the model accuracy. The same applies to *missing values* commonly occurring in many real-world data sets where data values might not be recorded for all variables.

Another commonly used approach is to replace the outlier or a missing value by an estimate of its true value. Typically, the average value is used. Although this approach introduces additional noise to the data, if the proportion of missing values is small, the effect on the modeling results might be tolerable.

2.2.2 Feature extraction

Feature extraction is one of the most important preprocessing steps of the machine learning process, having a major influence on the success of any subsequent modeling effort (Guyon & Elisseeff 2006). Feature extraction means creating new features as transformations or combinations of the original sensed data. Thus, feature extraction methods determine an appropriate linear or nonlinear subspace of the original data space (Jain *et al.* 2000). Features can be derived from either a single variable or a combination of variables in the raw data with the purpose of presenting the data in a form that best exposes its information content to the machine learning algorithm used. Feature extraction can also reduce the dimensionality or sensibility to noise of the original data set (Pyle 1999).

Even though the word *variable* is usually used in the context of statistical models and the term *feature* is commonly used for the inputs of a machine learning algorithm, we will use these terms somewhat interchangeably, depending on the context, in the following.

2.2.3 Normalization

Normalizing the data simply means scaling the data values to the range $[0,1]$. An easy way to do this is to use the *linear scaling transform*

$$x_{n_N} = \frac{x_n - \min(x_1 \ldots x_N)}{\max(x_1 \ldots x_N) - min(x_1 \ldots x_N)}, \tag{1}$$

where $x_1 \ldots x_N$ are all the values in the sample, x_n is the original instance value, and x_{n_N} is the normalized value (Pyle 1999). In a data set with several variables, each variable is normalized separately, which equalizes the magnitude of change in each of them. Normalization is a necessity for some modeling techniques and a convenience for others. For example, for techniques using the Euclidean distance, all variables should have the same scale for a fair comparison between them.

However, two kinds of problems might arise from normalization. First of all, the sample minimum and maximum values used in normalizing the data might be less extreme than the population minimum and maximum. Therefore, if these local extremes are used for normalizing future values, values outside the sample range will fall into a range wider than that intended $[0,1]$. The second type of problems arises if there are outliers in the sample. In this case, future values will be scaled to a range narrower than $[0,1]$, and the equalization of the magnitudes of change in the variables will be compromised.

In this work, we handled the first problem by either having a relatively large data set that was assumed to be fairly well representative of the whole population, or by expecting that when the model is applied in the future, for example for a participant not in our study, a new training set will be collected and the normalization will be done based on that. The second problem was prevented through either replacing values falling outside a known range of possible measurement values by the sample mean, through excluding a small proportion of observations at the extremes of the sample distribution, or by using the 5th and 95th percentiles of the sample values instead of the minimum and maximum in linear scaling so that these percentiles would meet the range $[0,1]$.

In spite of these shortcomings, normalization has the advantage that it preserves the original distribution of the variable.

2.2.4 Standardization

In *standardization*, each variable is transformed to have zero mean and unit variance. This can be done using estimates of mean and standard deviation calculated from the sample

$$x_{n_S} = \frac{x_n - \bar{x}}{\sigma}, \tag{2}$$

where x_n is the original instance value, $\bar{x}$ is the sample mean, σ is the sample standard deviation, and x_{n_S} is the standardized value.

Standardization is commonly used, especially when applying regression models (see Section 2.3.2) to the data, to improve the interpretability of explanatory variables and their relative impacts on the response variable, in particular. After standardization, the units of the regression coefficients are the same and they can be interpreted as the amount of change in the response variable when the explanatory variable increases by one standard deviation. However, the standardized coefficients depend on the sample variation. In spite of this, standardization can be useful in practice, especially as an automatic starting point to make coefficients roughly comparable (Gelman & Pardoe 2007). Standardization is also a linear transformation and it leaves the relative distances between observations intact.

Variables can also be *centered* around zero by subtracting the mean without dividing by standard deviation. Then the intercept term of a regression model is interpreted as the expected value of the response variable when the explanatory variables are set to their means. Otherwise, the intercept would be interpreted as the expected value of the response variable when the explanatory variables are set to 0, which may not be a realistic situation in practice.

2.3 Modeling

In this work, we have applied *predictive models* that aim to predict future values of a response variable based on input variables instead of *descriptive models* that help to understand the underlying process that generated the data. There are two kinds of predictive models depending on whether the response variable is categorical or real-valued. For a categorical response the modeling task is called *classification* whereas for modeling a real-valued response the term *regression* is used (Hand *et al.* 2001).

2.3.1 Classification

In classification, each observation, consisting of a p-dimensional vector $x = \{x_1, \ldots, x_p\}$ of input variables (also called features), is classified into one of m classes $c_1 \ldots c_m$. Classification methods can be divided into two categories, *parametric* and *non-parametric* classifiers. Parametric methods assume the data to originate from a certain distribution whose functional form is known. This simplifies the estimation and interpretation of models, but the resulting models may have relatively high bias because real-world data may not follow the assumed distribution (Hand *et al.* 2001). In addition, all of the commonly used parametric densities are unimodal, whereas many practical modeling tasks require multimodal densities (Duda *et al.* 2001). In non-parametric classification, on the other hand, few assumptions are made about the form of the underlying distribution so these methods can be used with arbitrary distributions.

In practice, it is often preferable to use simple classification rules if the data set available is small, to avoid *over-fitting* of the model to the data, whereas complex classifiers can be used more efficiently when a large data set is available (Raudys 2006). The same principle applies to regression models (Hand *et al.* 2001, Bishop 2006) which we will discuss later in the next section.

Parametric classifiers

The parametric methods introduced in the following, the quadratic discriminant analysis, linear discriminant analysis, Mahalanobis discrimination and Naïve Bayes classifier, model the class-conditional densities parametrically as multivariate normals. They are all applications of the *Bayes theorem* and classify an observation in the most probable of the classes. Bayes theorem states that the *a posteriori* probability of the class c_j, $P(c_j \mid x), j = 1 \ldots m$, can be estimated based on the class-conditional probability densities $p(x \mid c_j)$ and the class *a priori probabilities* $P(c_j)$

$$P(c_j \mid x) = \frac{p(x \mid c_j)P(c_j)}{p(x)}, \tag{3}$$

where $p(x)$ is the probability density function of x, which can be expressed as $p(x) = \sum_{j=1}^{m} p(x \mid c_j)P(c_j)$ (Theodoridis & Koutroumbas 2008). The density $p(x)$ is the same for all classes and it does not affect the classification decision.

The multivariate normal class-conditional densities have the form

$$p(x \mid c_j) = \frac{1}{(2\pi)^{p/2}|\Sigma_j|^{1/2}} \exp\{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)\}, \tag{4}$$

where μ_j is the mean and Σ_j is a $p \times p$-*covariance matrix* of the class c_j.

Quadratic discriminant analysis

By taking the natural logarithm of the numerator in Equation 3, we get the quadratic discriminant analysis (QDA) classification rule (Holmström *et al.* 1997)

$$g_{QDA}(x) = \operatorname*{argmax}_{c_j} \Big[-\frac{1}{2}\ln|\Sigma_j| - \frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j) + \ln P(c_j) \Big]. \tag{5}$$

The method is computationally relatively light, so it scales well to big data sets. However, it is not particularly reliable for large numbers of variables, as the dependence on the number of variables is quadratic (Hand *et al.* 2001).

Linear discriminant analysis

In linear discriminant analysis (LDA), the different classes are again assumed to have different mean vectors, but the covariances are now assumed to be equal. The LDA rule has the form (Holmström *et al.* 1997)

$$g_{LDA}(x) = \operatorname*{argmax}_{c_j} \Big[\mu_j^T \Sigma^{-1}(x - \frac{1}{2}\mu_j) + P(c_j) \Big]. \tag{6}$$

Even though QDA will likely fit the data better, the advantage of LDA is that it has fewer parameters to estimate. Both of these methods can often provide good classification results even if the normality assumption does not hold. However, for these methods to work, the number of observations in each class needs to exceed the number of input variables. Otherwise the covariance estimates are not positive definite and so cannot be inverted.

Mahalanobis discrimination

Mahalanobis discrimination (MD) is similar to the previous method, with the exception that the *a priori* probabilities of the classes are assumed to be identical, leading to the rule

$$g_{MD}(x) = \underset{c_j}{\operatorname{argmin}} \Big[(x - \mu_j)^T \Sigma^{-1} (x - \mu_j) \Big], \tag{7}$$

where the quantity to be minimized is called the *Mahalanobis distance* (Theodoridis & Koutroumbas 2008).

Naïve Bayes classifier

The Naïve Bayes classifier (NB) is based on the simplest assumption that the input variables x_i are conditionally independent given the class c_j. The classification rule is now reduced to (Mitchell 1997)

$$g_{NB}(x) = \underset{c_j}{\operatorname{argmin}} \Big[P(c_j) \prod_{i=1}^{p} p(x_i \mid c_j) \Big]. \tag{8}$$

In spite of the strong independence assumption, which is rarely true in real-world applications, the Naïve Bayes classifier is often quite effective. It reduces the p-dimensional multivariate problem to p univariate estimation problems where only the variances of the variables for each class need to be determined and not the entire covariance matrix. This is particularly helpful when the dimensionality of the input space is high. Since relatively few parameters need to be estimated, the Naïve Bayes classifier only requires a small amount of training data compared to the above methods with more relaxed assumptions. Even though the conditional independence assumption may lead to rather poor estimations of the class-conditional densities, the model may still perform well in practice because the decision boundaries can be insensitive to some of the details in the class-conditional densities (Bishop 2006). In addition, the decision surfaces produced by the Naïve Bayes classifier can in fact have complicated non-linear shapes that can fit quite complicated distributions (Hand *et al.* 2001). Because of the effectiveness, simplicity and robustness the Naïve Bayes classifier was selected as one of the 10 most influential data mining methods in 2008 (Wu *et al.* 2008).

Non-parametric classifiers

The non-parametric methods used in this work, the k-nearest neighbors classifier and the learning vector quantization, are based on modeling the classes using prototypes. In the case of learning vector quantization, the classification is performed according to the shortest distance to a prototype and in the case of k-nearest neighbors classifier according to the shortest distances to k prototypes, respectively.

k-nearest neighbors classifier

The prototypes of the *k-nearest neighbors classifier* (kNN) consist simply of the training vectors. All observations are assumed to correspond to points in the p-dimensional feature space and a new observation is classified to the class most frequently represented among its k nearest neighbors in the training set. Commonly, the Euclidean distance is used to define the nearest samples (Mitchell 1997). The value of the parameter k is usually chosen to be odd to avoid ties. In the case $k = 1$ the new observation is simply assigned to the same class as the nearest point from the training set.

The kNN method can also be seen as a way to estimate the *a posteriori* probabilities locally by

$$P(c_j \mid x) = \frac{p(x \mid c_j)P(c_j)}{p(x)} = \frac{k_j}{k}, \tag{9}$$

where k_j is the number of samples originating from class c_j (Bishop 2006). In fact, it has been shown that as the number of training samples N approaches infinity, the error rate of the *1*NN classifier is bounded above by twice the minimum achievable error rate of the optimal Bayes classifier (derived from Equation 3) that assumes the true class distributions to be known (Cover & Hart 1967). The asymptotic performance of kNN is even better since, as $k \to \infty$, the kNN tends to the Bayes classifier (Theodoridis & Koutroumbas 2008).

Unfortunately, these results do not hold in the finite sample case. In practice, we would like to use a large value for k in order to obtain a reliable estimate, since a small value would produce a classifier too sensitive to single data points. On the other hand, increasing k means that all the neighbors are not necessarily very close to the observation to be classified. The problem arises especially when the dimensionality p of the feature space is high: as p increases, while the number of samples N remains the same, the data becomes more and more sparse. This phenomenon is known as the so-called *curse of dimensionality* (Theodoridis & Koutroumbas 2008). As a compromise, we should choose a k that is a small fraction of the number of training samples N (Duda *et al.* 2001).

Another drawback associated with the *k*NN classifier is its computational complexity. As making a single classification decision requires visiting each of the training set samples to calculate the distance to them, the method becomes impractical for large data sets especially if the dimensionality is high. In addition, the entire training set needs to be stored in the memory.

Despite these problems, the *k*NN method is easy to implement, does not require training and may perform surprisingly well in some applications. Like the other simple classification method presented above, Naïve Bayes, *k*NN was also among the 10 most influential data mining algorithms selected in 2008 (Wu *et al.* 2008).

Learning vector quantization

In *learning vector quantization* (LVQ) (Kohonen *et al.* 1996), the prototypes, called codebook vectors, are composed of a more compact set of vectors based on the training samples. Once the number of codebook vectors has been decided on, they can be initialized, for example, by sampling from the set of training vectors correctly classified by the *k*NN method, or by using the *self-organizing map* (SOM) algorithm (Kohonen *et al.* 1996). Then, the locations of the codebook vectors are updated iteratively using each training sample in turn. The principle of LVQ is that the training samples attract codebook vectors of their own class and reject prototypes of other classes.

The codebook vectors are used in classification according to the *1*NN rule. An observation x is classified to the same class as the closest codebook vector m_i in the sense of Euclidean distance. The index of the nearest prototype can be defined as

$$c = \underset{i}{\operatorname{argmin}}\{||x - m_i||\}. \tag{10}$$

The basic LVQ algorithm (LVQ1) is defined by the following algorithm, where $x(t)$ is an input vector and $m_i(t)$ represents sequential values of the m_i at discrete time steps $t = 0, 1, 2, \ldots,$

$$\begin{aligned} m_c(t+1) &= m_c(t) + \alpha(t)[x(t) - m_c(t)], \\ &\qquad \text{if } x \text{ and } m_c \text{ belong to the same class,} \\ m_c(t+1) &= m_c(t) - \alpha(t)[x(t) - m_c(t)], \\ &\qquad \text{if } x \text{ and } m_c \text{ belong to different classes,} \\ m_i(t+1) &= m_i(t) \text{ for } i \neq c. \end{aligned} \tag{11}$$

The learning rate parameter $\alpha(t), 0 < \alpha(t) < 1$, is usually made to decrease monotonically in time. The class borders defined by the LVQ algorithm are piecewise linear.

There are several variations of the LVQ algorithm that differ in the way the prototype vectors are updated. In the *optimized learning rate* LVQ1 (OLVQ1), the learning rate parameter $\alpha_i(t)$ is chosen individually for each codebook vector m_i. In the LVQ3 algorithm, the two closest codebook vectors are updated simultaneously. Two variants of the LVQ algorithm based on cost functions, the *generalized learning vector quantization* (GLVQ) (Sato & Yamada 1996) and *soft learning vector quantization* (SLVQ) (Seo & Obermayer 2003) improve the classification ability. GLVQ satisfies the convergence condition of the rule and SLVQ extends the LVQ family to different distance measures.

Compared with the other prototype classifier introduced above, *k*NN, the limitation of LVQ is the large number of steps needed. Where in *k*NN there is only one parameter value to decide on, *i.e.* the number of nearest neighbors k, and no training is needed, in LVQ the classification result and the time needed for learning depend on several factors. First, the user needs to fix the number of codebook vectors and choose a method to initialize them. Then, the specific LVQ algorithms to be used, the value of the learning rate and a stopping criterion for learning need to be decided on (Kohonen *et al.* 1996).

The advantage of LVQ over *k*NN is that not the whole training set needs to be stored for reference, but a smaller number of codebook vectors. This is especially the case since good approximations of the class-conditional densities $p(x \mid c_j)$ are not needed everywhere, but only at the class borders (Kohonen *et al.* 1996).

2.3.2 Regression

The goal of regression is to find a functional description for data that can be used to predict values of the response variable for new input. The simplest and most popular form of regression is *linear regression*, in which the function is linear in the input variables (Hand *et al.* 2001). The linear regression model predicts values of the response variable y as a linear combination of the explanatory variables $x_i, 1 \leq i \leq p$. This can be presented for all the observations simultaneously with the matrix form

$$Y = X\beta + \varepsilon, \tag{12}$$

where Y is the N-dimensional vector of responses, X is a $(N \times p)$-dimensional matrix of the explanatory variables, β is the vector of regression coefficients to be estimated and ε is a random error. The errors for different observations are assumed to be independent and identically distributed, and follow the normal distribution, denoted $\varepsilon \overset{i.i.d.}{\sim} N(0, \sigma^2 I)$. The first column of X contains a vector of 1s to include an intercept term in the model. Commonly, the coefficients of the model are estimated by minimizing the sum of squared errors, a procedure known as the *least squares method*, that under the assumption of independent and identically distributed observations coincides with the *maximum likelihood method* (Borovkov 1999).

Hand *et al.* (2001) explain the popularity of this simple method because it is very easy to both compute and understand. It also often performs well even when the true relationship between the input and response variables is not linear. On the other hand, they point out that estimation of the coefficients becomes difficult if the sample size N is small, or if the input variables are exactly or almost linearly dependent.

Linear mixed model

The *linear mixed model* is particularly suited for modeling data originating from different statistical units, such as test subjects, with repeated measurements. Unlike the linear regression model, which assumes that observations are independent and identically distributed, the linear mixed model assumes two sources of variation: *subject-specific* and *population-specific*. The vector of repeated measurements of each subject is assumed to follow a linear regression model where some of the regression parameters are common for all the subjects, and others differ between subjects.

In general, a linear mixed model has the form (Verbeke & Molenberghs 2000)

$$\begin{aligned} Y_s &= X_s\beta + Z_s b_s + \varepsilon_s, \\ b_s &\overset{i.i.d.}{\sim} N(0, D), \\ \varepsilon_s &\overset{i.i.d.}{\sim} N(0, \Sigma_s), \quad 1 \leq s \leq S, \end{aligned} \tag{13}$$

where S is the number of subjects, Y_s is the N_s-dimensional response vector for subject s, X_s and Z_s are $(N_s \times p)$ and $(N_s \times q)$ dimensional fixed matrices of explanatory variables,

β is a p-dimensional unknown fixed-effects parameter vector, b_s is the q-dimensional vector of random effects, and ε is an N_s-dimensional vector containing the error components. D is a general symmetric $(q \times q)$ covariance matrix and Σ_s is a $(N_s \times N_s)$ covariance matrix where the set of unknown parameters in Σ_s do not depend upon s.

If the only subject-specific term in the model is set to be the intercept and the regression coefficients are otherwise the same for all subjects, we get the *random-intercepts model*. For more information on the linear mixed model, see (Verbeke & Molenberghs 2000).

2.4 Model selection

In this work, we have used methods relating to two different aspects of model selection: *validation* and *dimensionality reduction*, that we will discuss in this section.

As the goal of model selection is to find a model that best describes the data or predicts future values as accurately as possible, the natural first step is to choose a *criterion* or *score function* to compare model candidates. In classification, an obvious choice is the *misclassification rate*. For regression models, the *log-likelihood function* is commonly used (Hand *et al.* 2001).

2.4.1 Validation

To ensure the best predictive performance of a model on new data, and to make sure no over-fitting has taken place, a data set separate of the *training set*, called *validation set*, needs to be used to estimate the model generalization performance. The simplest approach for this is to split the available data into two parts and to use the first one for adjusting the model parameters and the second to estimate the generalization error. Generally, a smaller portion of the data is used for validation than for training (Duda *et al.* 2001). The same approach can be used to compare a range of different models to find the most appropriate one for a given application.

Over-fitting in model selection is likely to be most severe when the sample of data is small and the collection of models to choose from is relatively large (Cawley & Talbot 2010). If the model design is iterated several times, or if the set of candidate models from which the final model is selected is large, it is necessary to keep aside a third *test set* on which the performance of the final selected model is evaluated. Otherwise, if the model performing best on the validation set is selected, the performance estimate of the final model will be optimistically biased towards this set (Hand *et al.* 2001, Bishop 2006).

Many times, however, the data set size is so limited that we wish to use as much of it as possible for training the model. In this case, a generalization of the above method, called *M-fold cross-validation*, can be adopted. The training set is divided into *M* sets (usually of equal size), and the model is trained *M* times, each time holding out a different set for validation. The overall performance of the model is then calculated as an average of these *M* errors (Duda *et al.* 2001). If the number of hold out sets *M* is selected to be equal to the number of observations in the data set *N*, the method is called *leave-one-out* cross-validation.

One drawback of the cross-validation approach is that the number of training runs needed is increased by a factor of *M* (Bishop 2006). Another concern is that, for small data sets, because of variation across the data sets, the variance of the cross-validation error can be unreasonably high, and thereby needs to be monitored in practice (Hand *et al.* 2001).

In the case of cross-validation, as well, it is important to use a separate test set or holdout sets generated by an external loop of cross-validation that were never used during the model selection process (Moore & Lee 1994, Kohavi & Sommerfield 1995, Reunanen 2003). Otherwise, an intensive use of cross validation may produce a deceptively good lowest-error model, in a manner similar to over-fitting of data (Schaffer 1993, Moore & Lee 1994). Cawley & Talbot (2010) recommend to always use multiple partitions of the data to form training, validation and test sets, as the sampling of data for a single partition might arbitrarily favor one model over another.

In spite of the recurrent mention of the importance of a separate test set in evaluating the selected model, a common mistake, even among experienced machine learning researchers, pointed out by Kohavi & Sommerfield (1995), Reunanen (2003), Fiebrink & Fujinaga (2006), and Smialowski *et al.* (2010), is to present the training set or cross-validation estimate as the final performance estimate of a model, thus achieving overly optimistic results.

2.4.2 Dimensionality reduction

As we have seen earlier in this chapter, many of the presented methods suffer from high dimensionality. In addition to model accuracy and robustness, a high number of features also negatively influences the computational time of the algorithms, and possibly increases the costs of data collection and storage. Models with fewer features are often also easier to interpret. There are two approaches to dimensionality reduction that we shall discuss in this section: *feature subset selection*, and *transforming the feature space*.

Feature selection

The goal of feature selection is to select the best subset of the original feature set. The feature selection process can be started by inspecting each feature individually to discard features that do not carry enough information related to the current modeling problem. However, since there may be dependencies between the features, selecting the individually best features may not lead to the best possible result. With a small number of features it may be possible to search through all possible feature combinations. In practice, this is rarely the case, and we need a more sophisticated search strategy to go through the space of possible feature sets.

Sequential selection methods

The *sequential forward selection* (SFS) (Whitney 1971) is a simple bottom-up search procedure in which one feature at a time is added to the current feature set. At each stage, the feature to be included is selected from the set of remaining available features, so that the new extended feature set yields a maximum value of the criterion function used (Devijver & Kittler 1982). The sequential backward selection (SBS) method is the top-down counterpart of the SFS algorithm. The advantage of these algorithms is their easy application and relatively short calculation time.

Sequential floating selection methods

The floating forward and backward feature selection methods, *sequential forward floating selection* (SFFS) and *sequential backward floating selection* (SBFS), introduced by Pudil *et al.* (1994), are based on the *plus l - take away r* method (Stearns 1976), in which the feature set is alternately enlarged by *l* features using the SFS method, and reduced by discarding *r* features applying the SBS algorithm. In the floating selection methods, however, the number of forward and backward steps is dynamically controlled instead of being fixed in advance. The conditional inclusion and exclusion of features is controlled by the value of the criterion function. In the bottom-up algorithm, SFFS, after each forward step, a number of backward steps are applied as long as the resulting subsets yield better values for the criterion function than the previously evaluated ones of the same dimension. In the top-down counterpart, SBFS, an exclusion of a feature is followed by a series of successive conditional inclusions if an improvement to the previous sets can be made. The feature to be included into the current feature set or excluded from it is always the one that improves the set most, or degrades the value of the criterion function least (Pudil *et al.* 1994).

Compared to the basic sequential feature selection methods, the main advantage of the floating methods is that the resulting feature sets of different dimensions are not necessarily nested, as in the case of the SFS and SBS methods. The floating methods are able to correct erroneous decisions made at the previous steps of the algorithm. Therefore, these methods provide a close to optimal solution to the problem of feature subset selection. Because of this characteristic, they are also applicable to problems involving non-monotonic feature selection criterion functions (Pudil *et al.* 1994). In addition, even though the floating feature selection methods are only nearly optimal, they are much faster than the optimal but computationally exhaustive *branch and bound algorithm* (Narendra & Fukunaga 1977).

After the discussion of over-fitting in feature selection by Wolpert (1992), Schaffer (1993) and Kohavi & Sommerfield (1995), Ambroise & McLachlan (2002) and Reunanen (2003) pointed out a methodological flaw in the comparisons claiming to show the superiority of the floating methods compared to the basic sequential methods (*e.g.* Kudo & Sklansky (2000)). In many studies, the reported results were based on cross-validation accuracy in the training set only where over-fitting might have happened and no separate validation set was used. According to Reunanen's (2003) results, the basic algorithms may give equally good results in significantly less time, although the floating methods' capability to avoid nesting of the subsets can in some cases be useful. Reunanen (2004) even showed that, when evaluated properly, feature selection is often actually ineffective at improving classification accuracy.

Nevertheless, the floating methods are still widely used and many researchers (e.g. Peng *et al.* (2010)) consider them as one of the best alternatives for feature selection. This, however, does not contradict with Reunanen's (2003) findings as they did not criticize the methods themselves, only the implementation of the studies aiming to show their superiority.

n best features selection

The *n best features selection* method (*n*Best) simply means selection of the *n* individually best features in the sense of maximizing the criterion function. It is the simplest alternative for feature subset selection, but also the most unreliable since the features selected may correlate with each other. However, the use of simple search algorithms that are less prone to over-fitting might be justifiable in the case that very little data is available (Reunanen 2003).

Feature over-selection

Later, Raudys (2006) showed that *feature over-selection* can decrease the performance of a classifier, especially when the validation set used is very small or when a very large number of feature subsets is considered. Consequently, the emphasis in feature selection algorithm comparisons has recently shifted from searching for the most optimal feature subset with respect to some criterion function to aiming at a feature subset with the best generalization performance, *i.e.* the ability to perform well on previously unseen data. For example, Saari *et al.* (2011) considered the generalizability and simplicity of the obtained models as criteria for feature selection in classification of mood in music.

Cawley & Talbot (2010) also noted that if the difference in performance between two feature selection algorithms is smaller than the variation in performance due to the random sampling of the data, it is unlikely to be of practical significance. In that case, a greater improvement in performance would be obtained by further data collection than by selection of the optimal classifier.

Transforming the feature space

Another approach to dimensionality reduction is to transform the original feature set onto a lower dimensional subspace. The basic idea is to transform the original p-dimensional input vectors $x = \{x_1, \ldots, x_p\}$ into p'-dimensional vectors $z = \{z_1, \ldots, z_{p'}\}$, where typically p' is much smaller than p. The z variables are defined as functions of the x variables, and the transformation is chosen in some sense to produce the best set of p' variables for the task at hand (Hand *et al.* 2001).

Principal component analysis

In *principal component analysis* (PCA) (Hotelling 1933) the original data is projected onto a lower dimensional space, and the projection is chosen so as to maximize the variance of the projected data. In practice, the transform is accomplished by projecting the original data linearly onto an orthogonal basis defined by the p' eigenvectors $u_1, \ldots, u_{p'}$ of the data covariance matrix Σ corresponding to the p' largest eigenvalues $\lambda_1, \ldots, \lambda_{p'}$. Thus, the resulting features are mutually uncorrelated.

Since the eigenvalues are equal to the variances of the transformed features, the total variance of the projected data can be calculated as $\sum_j \lambda_j$, where λ_j is the jth eigenvalue. In addition, this projection minimizes the *mean squared error* (MSE) compared with any other approximation of the original p-dimensional feature vector x by an p'-dimensional vector (Pearson 1901).

The squared error of approximating the original data can be expressed as (Hand *et al.* 2001)

$$\frac{\sum_{j=p'+1}^{p} \lambda_j}{\sum_{l=1}^{p} \lambda_l}. \tag{14}$$

Thus, one approach for choosing the number of principal components to be used is to increase p' until the squared error is smaller than some predefined threshold. Another way is to plot the eigenvalues in descending order by their magnitude, to demonstrate the amount of variance explained by each of them, and to choose the ones that have the largest values. This is particularly helpful if there is a sudden fall towards zero somewhere in the plot (Hand *et al.* 2001).

PCA is a powerful tool with excellent information packing properties. However, it does not always lead to maximal class separability in the projected feature space since the dimensionality reduction is not optimized with respect to this property (Theodoridis & Koutroumbas 2008). It is also not invariant under rescalings of the original variables, so the data is typically normalized before applying PCA if different variables are measured in different units (Hand *et al.* 2001).

2.5 Special considerations on time series modeling

In this thesis, we have chosen to use the term *time series data* to emphasize the time-dependent nature of the sensor data we used. Elsewhere, other terms, such as *streaming data* or *sequential data* have been used in similar but not identical contexts. *Data stream mining* is usually understood to be dealing with very high fluctuating data rates with underlying distributions changing over time (Dietterich 2002). Sequential data, on the other hand, is a broad concept also covering time series data, but also other types of sequences not ordered by time stamps (Xing *et al.* 2010).

The general considerations and methods presented so far in this chapter apply mostly to time series data as well. However, the temporal dependency of time series data adds some additional aspects to the modeling process, and sets some restrictions on the direct application of the methods in practical time series machine learning problems.

Time series data is typically high-dimensional, and characterized by noise and trends that are removed using preprocessing methods. In this work, we have used the Reinsch algorithm (Reinsch 1967, 1971) and convolution with a Bartlett window (Prabhu 2013) to smoothen, and differencing (Chatfield 2003) to remove trends.

The purpose of feature extraction is to compress the time series, keeping only the important information. It can be based on *descriptive statistics* (Kugiumtzis & Tsimpiris 2010), *time series analysis* (*e.g.* auto-correlation, autoregressive models) (Chatfield 2003, Hyndman & Athanasopoulos 2014), or *frequency domain* features (*e.g.* discrete wavelet transform, discrete Fourier transform, energy on frequency bands) (Agrawal *et al.* 1993, Mörchen 2003, Gevins & Rémond 1987). Typically, features are calculated on subsequent segments of time series observations, called *windows*. A special type of feature, past observations (*lags*), is important for time series prediction models (Chatfield 2003, Hyndman & Athanasopoulos 2014).

When time series data is collected from different statistical units (for example test subjects), both the *within-subject variation* and the *between-subject variation* in the data needs to be taken into consideration when selecting an appropriate model type. We can expect the repeated responses from the same subject to be more similar than the responses across different subjects (Fitzmaurice *et al.* 2011). This means that a model describing variation in one subject's data may not be generalizable to the whole population, and on the other hand, a model fitted to the data from the entire population of subjects may not explain individual variation in a single subject's data. In the context of linear regression, we have used a particular model suitable for this purpose, the linear mixed model (discussed earlier in Section 2.3.2) in this work.

Also when dividing a data set for training, testing and validation, the temporal dependency of time series data needs to be considered. If a data set collected during one session is randomly divided into three parts, or random samples are held out for cross-validation, one ends up training on data that is temporally fairly close, or even adjacent to test data. Because the adjacent observations are not independent, this produces significantly overestimated estimates of the model performance. In statistics, this phenomenon is well-known (Hart & Wehrly 1986, Burman *et al.* 1994, Arlot & Celisse 2010). According to Arlot & Celisse's (2010) review of cross-validation procedures for model selection, the most commonly used approach for applying cross-validation for dependent observations is to use a distance threshold from which the observations are independent. This method, called *modified cross-validation* was introduced by Chu & Marron (1991) and applied for nonparametric regression. In the field of machine learning, the effect of dependent observations on validation performance is less commonly acknowledged. In part of this work, we have used a block study design adapted from the *block cross-validation* scheme (Grimes *et al.* 2008), where entire contiguous sequences of observations are used for training, model selection and validation. Even better would be to use distinct data sets collected on different occasions.

3 Results

This chapter presents the main results of our case studies, where we used model selection methods to optimize model performance. The findings of this thesis are drawn from studies in three different application areas. Section 3.1 introduces our study on improving the quality of resistance spot welding processes (Publications I and II). Section 3.2 presents our results on human exercise energy expenditure estimation, based on measurements from a wrist-worn acceleration sensor (Publication III), and Section 3.3 discusses modeling the human cognitive load, based on psycho-physiological measurements (Publications IV and V). Even though the scientific contributions of the individual publications lie in their application areas, we can use these studies as examples when considering different aspects of the modeling process. Section 3.4 summarizes the contributions of the publications from the point of view of improving model performance, and the work is discussed as a whole in Chapter 4.

Table 1 summarizes the data characteristics and model requirements of the three case studies with the type of sensor data available, desired model outcome, type of modeling task (classification / regression), as well as the temporal and computational requirements of the model.

Table 1. Data characteristics and model requirements of the three case studies.

Study	Data	Model outcome	Type	Temporal requirements	Computational resources
Resistance spot welding	Current and voltage signals	Most similar previous welding process	C	Offline	Not limited
Exercise energy expenditure	Acceleration signals	Energy expenditure estimate in VO_2 / kg	R	Online	Limited
Cognitive load	Psycho-physiological sensor streams	Cognitive load estimate (low/high)	C	Online	Not limited [1]

[1] Possibly limited in a future implementation (*e.g.* a mobile device)

3.1 Resistance spot welding

Resistance spot welding (RSW) is one of the most important methods for joining metal objects. In the automotive industry, a typical passenger car requires approximately 4000 welding joints and the durability of the vehicle is dependent on the quality of the welding on the body (Wylie *et al.* 2010). In RSW, two or more metal sheets are joined together by passing an electrical current through them. The current is conducted through two electrodes pressed against the metal surfaces to hold the parts to be welded tightly together. The heat produced by the flowing current melts the metals, and a welding spot is formed. The amount of current, pressure and time are all carefully controlled and matched to the type and thickness of the material.

In general, the bigger the diameter of the welding nugget, the firmer is the joint. However, this dimension does not only depend on the welding parameters applied, but also on the profile of the electrode tip that wears out in time. Some other factors, such as faults and embrittlement in the welding joint, also affect its strength. The most reliable and commonly used method to verify the quality of an RSW joint is to tear the welded parts apart after cooling to measure the spot diameter. However, the welding joint is thereby destroyed. Some non-destructive methods for estimating the spot diameter have also been used (e.g. radiographic and ultrasonic weld inspection (Anderson 2001) and primary circuit dynamic resistance monitoring (Cho & Rhee 2002)), but the challenge has been to find a real-time, non-destructive method for online use in production lines.

More recently Cullen *et al.* (2008) introduced a method for online real-time non-destructive quality control of RSW that is based on ultrasonic monitoring and neural network modeling, which evaluates every weld as it is formed. Similarly, El Ouafi *et al.* (2012) presented a dynamic resistance based model for online quality control of RSW.

3.1.1 Process classification

The objective of this study was to use information collected from previous welding processes to reduce the set-up time of a new process. We compared the characteristics of a sample from a new welding process to information collected from previous processes to find a process that would match the new process as closely as possible. Then, the process parameters and quality control methods proven to lead to high-quality welding joints for the previous process could be applied for the new process.

In Publication I, we searched for suitable features calculated from the welding signals to present their characteristics and looked for a classifier that would give the best results in classifying welding processes. In Publication II, we further explored the feature space to eliminate features with less classification-relevant information to speed up the classification, and to improve the classification accuracy.

3.1.2 Data

The data sets we used in this study were supplied by two welding equipment manufacturers: Harms+Wende GmbH & Co.KG (HWH) and Stanzbiegetechnik GES.M.B.H. (SBT). Sets of welding experiments conducted with different welding machines, materials and thicknesses of the objects to be welded were called welding processes. There were altogether 20 processes, of which 11 were from HWH, and 9 from SBT. The data set comprised of a total of 3879 welding experiments. Each of the observations contained measurements of current and voltage recorded during an RSW event.

3.1.3 Features

We extracted 12 geometrical and 15 statistical features from the two signal curves relating to a single welding experiment (thus totaling 54 feature values). The geometrical features were chosen to locate the transition points of the curves as precisely as possible. The statistical features included the median of the signal and the means of the signal values calculated on four intervals based on the transition points. In addition, the signal curve was divided into ten intervals of equal length, and the means of the signal values within these intervals were used as features.

We then formed eight different combinations of the original features to be tested with different classifiers. The first feature set contained all the features, while the second consisted of only the ten means. Since the number of features was rather high, and we did not known if all of them contained information relevant to the classification, we then compressed the two feature sets using principal component analysis (PCA). Finally, the last four feature sets were obtained by standardizing each of the previous sets to have a mean of zero and a standard deviation of one.

3.1.4 Classifier selection

In Publication I, we tested five different classifiers to classify the welding processes. Since the distribution of the data was unknown, we chose both parametric and non-parametric methods. The classifiers we used were quadratic discriminant analysis (QDA), linear discriminant analysis (LDA), Mahalanobis discrimination (MD), the *k*-nearest neighbor classifier (*k*NN) and learning vector quantization (LVQ).

In order to evaluate the classifiers, we divided the data into training and test data sets, which consisted of 2/3 and 1/3 of the data, respectively. The training data set was used to train each of the classifiers, and the test set was used to evaluate their performance.

Before the actual classification, we searched for suitable parameter values for the *k*NN and LVQ classifiers using 10-fold cross-validation on the training set. Figure 1 presents the cross-validation results for *k*NN, where the best value for the parameter k and the number of principal components to be used was searched for. The surface plot demonstrates the results for the feature set consisting of all the features. As we can see, the classification accuracy does not notably improve after the inclusion of the fifth principal component. Likewise, the value 3 for the parameter k yields good results in the classification. The results for the other feature sets were similar.

Figure 2 shows how an increase in the number of LVQ prototype vectors, called codebooks, affects the accuracy of classification. We selected the parameter value 2200 because it seems to yield good classification results for all the feature sets. As advised by (Kohonen *et al.* 1996), the LVQ codebooks were initialized using the *3*NN algorithm, after which the OLVQ1 and LVQ3 algorithms were used to train the model at a learning rate of 0.05.

We tested the five classifiers on the eight feature sets, and the results for the test data are shown in Table 1. The percentages in the cells indicate the ratios of correctly classified processes; the cells left empty indicate invalid classifier - feature set combinations due to high feature correlation.

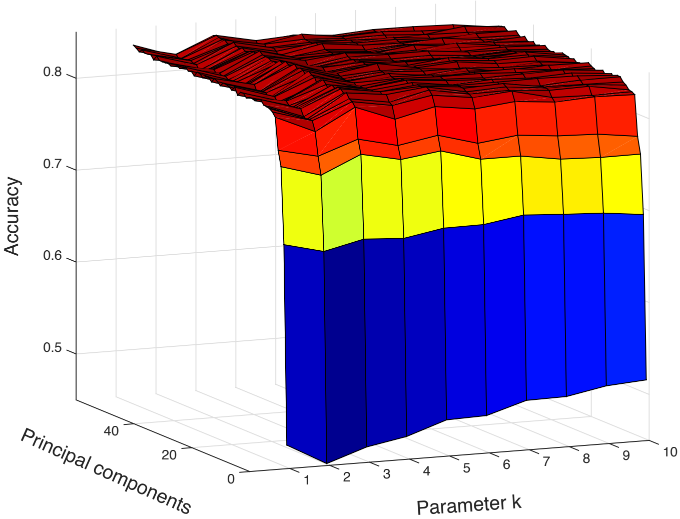


Fig 1. A surface plot of the results of tenfold cross-validation of the parameter *k* and the number of principal components used. Reprinted with permission from Publication I ©2005 Springer.

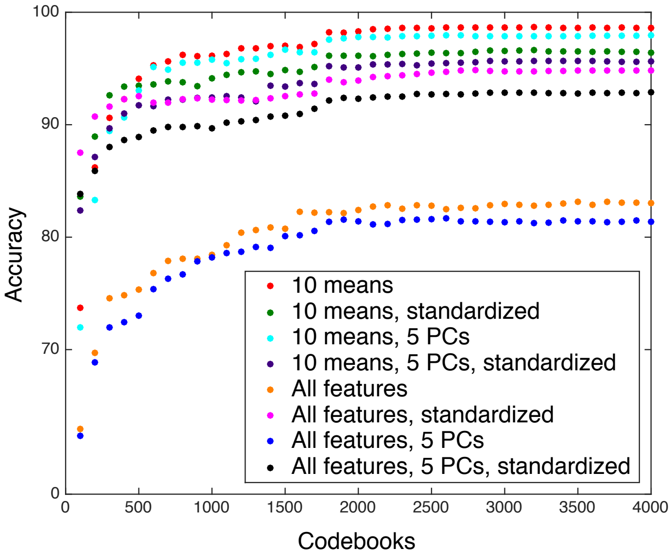


Fig 2. Results of tenfold cross-validation of the number of LVQ codebook vectors for the different feature sets. Reprinted with permission from Publication I ©2005 Springer.

Table 2. Classification accuracies for 20 welding processes with different classifiers and feature sets. Revised from Publication I ©2005 Springer.

	LDA	QDA	MD	*3*NN	LVQ
All features	93.0			84.1	84.5
All features, 5 PCs	62.5	75.2	72.4	83.2	82.5
All features, standardized	93.0			94.7	94.9
All features, standardized, 5 PCs	71.1	85.5	86.3	93.5	92.4
10 means	90.9	96.4	97.1	98.5	98.1
10 means, 5 PCs	82.1	94.3	94.4	97.8	97.1
10 means, standardized	90.9	96.4	97.1	95.4	96.1
10 means, standardized, 5 PCs	76.2	89.3	88.3	94.6	94.1

Classification accuracy was dependent on both the feature set and the classifier used. QDA performed better than LDA, indicating that the data rather supported quadratic than linear decision boundaries. The results for the MD method were approximately equal to QDA. However, none of these classifiers compared with the two non-parametric prototype classifiers, *k*NN and LVQ, that gave the best classification results and performed approximately equally well. We found the *k*NN classifier to be most suitable for this study due to its easy implementation in contrast to LVQ. The *3*NN classifier with the 10 signal means as features was the best classifier – feature set combination, with a classification accuracy of over 98%.

3.1.5 Feature selection

In Publication II, we then tested several feature selection methods on the original feature set to consider the usefulness of the individual features, and to further reduce the dimension of the feature space. The algorithms we tested were the sequential forward selection (SFS), sequential backward selection (SBS), sequential forward floating selection (SFFS), sequential backward floating selection (SBFS), and the *n* best features selection (*n*Best). We used the classification accuracy of the *3*NN classifier as the criterion function for the feature selection. We applied the feature selection methods to both the original and the standardized feature set. The best classification results for each method are presented in Table 3 a) and b).

Table 3. Classification results of the best feature subsets for five feature selection methods together with the number of features included in each set. a) Original feature data b) Standardized feature data. Revised from Publication II ©2008 Springer.

a)

	SFS	SBS	SFFS	SBFS	nBEST
Classification accuracy	98.9	84.7	99.3	84.8	95.3
Number of features used	6	7	11	7	10

b)

	SFS	SBS	SFFS	SBFS	nBEST
Classification accuracy	98.3	97.1	98.5	97.5	95.1
Number of features used	18	17	19	29	52

The best results were obtained with the forward methods applied to the original feature set. The SFFS method produced the best feature subset, with 11 features and a classification accuracy of 99.3%. The SFS method provided a set of 6 features with a comparable classification accuracy of 98.9%. The subsets selected from the standardized feature set are notably larger than the subsets of the set of original features, and only the backward methods, SBS and SBFS, seem to yield better feature subsets when applied to the standardized data. The fact that standardization seemed to weaken the classification results implies that the features calculated from the current signal, which originally had the wider measurement range, contained more classification-relevant information than the features calculated from the voltage signals. The standardization equalized the magnitude of the two feature types and hence obscured some of the information relevant to the classification.

3.2 Exercise energy expenditure estimation

Daily physical activity has important and wide-ranging health benefits. For adults, regular physical activity has been shown to reduce the risk of chronic diseases such as heart disease, type 2 diabetes, and some cancers, and also enhance and preserve function with age (Blair 2009). For children, in addition to preventing obesity, promoting a physically active lifestyle reduces the risk of getting these diseases later in life (Froberg & Andersen 2010). Objective and reliable assessment of physical activity is essential to evaluate these health benefits, to give recommendations on the amount of daily exercise and to track adherence to these recommendations. Energy expenditure (EE) caused by physical activity is commonly accepted as the standard reference of physical activity

(LaPorte *et al.* 1985). It can be reliably measured from a person's oxygen consumption. However, measurement of oxygen consumption requires the use of a breath gas analyzer (indirect calorimetry), and is therefore impractical and not feasible under free-living conditions.

3.2.1 Modeling physical activity based on body-acceleration

Research on modeling physical activity in different activities based on acceleration data has expanded over the past two decades (Bouten *et al.* 1994, Troiano 2006). Measurements done with body-mounted accelerometers are widely used in determining the frequency and intensity of movements during physical activity (Bouten *et al.* 1997). Estimation of EE based on acceleration measurements has become a widely discussed problem approached in various studies. However, widely accepted, precise and reliable methods for estimating physical activity based on acceleration data have not been found.

In many studies, regression methods have been applied to accelerometer counts, obtained by integrating the accelerometer signal (Chen & Bassett 2005), and oxygen consumption simultaneously measured to determine the relationship between the two measures and to define an equation to predict EE from acceleration (Albinali *et al.* 2010, Troiano 2006). Some of the most widely used regression models are the Freedson (Freedson *et al.* 1998), Swartz (Swartz & Strath 2000), and Hendelman (Hendelman *et al.* 2000) equations. Also neural networks have been used to map the activity count data (Staudenmayer *et al.* 2009) or raw acceleration data (Rothney *et al.* 2007) to EE. In most studies, one acceleration sensor has been used on the participants waist (Troiano 2006) or hip (Rothney *et al.* 2007), but some studies have used additional sensors placed on other body parts such as the thigh, upper arm or wrist (Albinali *et al.* 2010, Swartz & Strath 2000).

As Rothney *et al.* (2007) mention, one of the biggest challenges in modeling EE based on accelerometer data is the large deviation between subjects. Because of different personal characteristics, the same metabolic costs may not result from identical accelerations. In Publication III, we introduced a new method of modeling exercise EE based on wrist-acceleration using the linear mixed model. The advantage of the linear mixed model is that it models both between-subject and within-subject variation. Therefore, individual differences in EE can be accounted for.

3.2.2 Data

The data we used in this study was provided by Polar Electro Oy. It was collected from ten healthy participants (8 men and 2 women) whose characteristics are shown in Table 7. The participants performed four different activities: walking, running, Nordic walking and bicycling for 10 minutes each. These activities were selected because they constitute a considerable proportion of the incidental exercise performed in daily life; that is, the exercise we get doing daily activities. In addition, these activities were suitable for showing the functionality of the algorithm proposed, since they differ in the amount of movement of different body parts.

Table 4. Physical characteristics of the subjects (mean, standard deviation, range). Reprinted with permission from Publication III ©2008 IEEE.

	Men (n=8)	Women (n=2)	All (n=10)
Age (years)	29.8 ± 4.7 (22-37)	24.5 ± 4.9 (21-28)	28.7 ± 4.9 (21-37)
Height (cm)	181 ± 5.6 (169-188)	164 ± 1.4 (163-165)	178 ± 8.8 (163-188)
Body mass (kg)	82.6 ± 13.1 (62-104)	56.5 ± 7.8 (51-62)	77.4 ± 16.2 (51-104)
BMI (kg/m^2)	25.1 ± 3.4 (21.6-31.1)	21.0 ± 2.5 (19.2-22.8)	24.3 ± 3.5 (19.2-31.1)

While performing the activities, the participants were asked to wear a biaxial accelerometer on their left wrist, as well as a portable breath gas analyzer (Cosmed K4b2) that uses a face mask. The sampling rate of the accelerometer was 100 Hz and the breath gas exchange measurement system recorded the concentration of expiration gases at intervals of 15 seconds. The weight of the subject was taken into consideration by dividing the oxygen consumption values by the weight of the subject.

Although all ten participants performed the four activities, the measurement data could not be recorded from all the activities due to technical problems with the accelerometer data logger. The bicycling data set contains measurements from nine participants, while the walking and Nordic walking data is available from seven participants, and the running data set has measurements from six participants.

3.2.3 Features

Instead of the accelerometer counts commonly used in other studies, we used the variance of the raw acceleration signal as a feature as we believed it would better preserve the intensity of acceleration. Since the sampling frequency of the acceleration measurement was higher than that of the expiration gases, we calculated the variances at 15-second intervals on the two acceleration signals. Other features we used in the models were the height and a logarithmic transformation of the height of the subject, and the product of the two variance values. In addition to the acceleration measurements made simultaneously with the oxygen consumption measurement, we used the five preceding variance values to take into account the influence of acceleration from the 90 seconds preceding the oxygen consumption measurement.

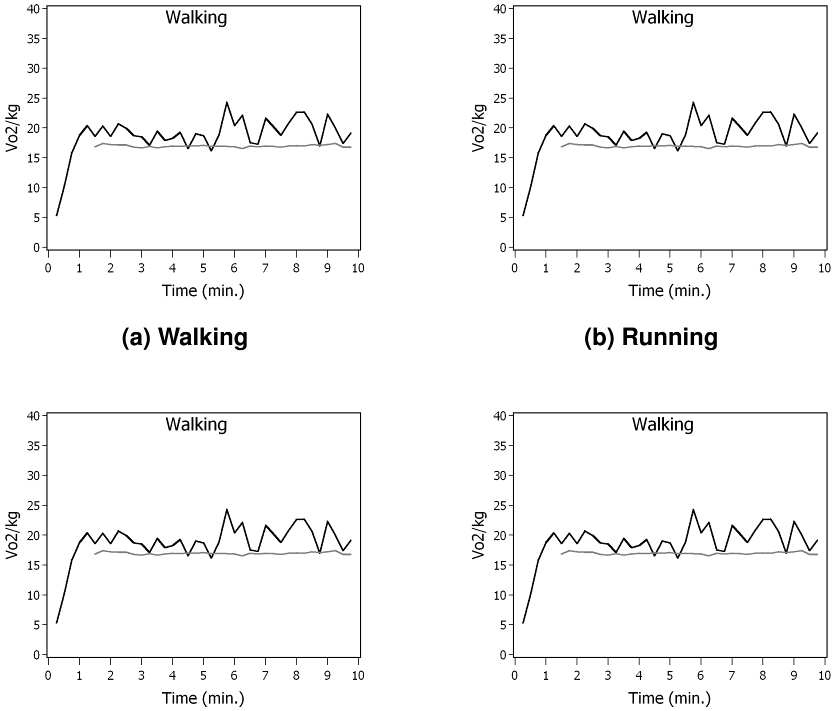
3.2.4 Model type and feature selection

We modeled the oxygen consumption of the participant using the linear mixed model that models both between-subject and within-subject variance. This type of model was chosen to solve the estimation problem, since it is very suitable for modeling time series data containing repeated measurements from several participants which are correlated with each other.

An overall intercept term was fitted to model the average level of oxygen consumption across all participants, and subject-specific intercepts to model the individual ground level of each participant. This is the so-called *random-intercepts model* (Verbeke & Molenberghs 2000), where the regression coefficients are the same for all participants. When the model is applied to measurements outside the training data set, no information about the subject-specific level of oxygen consumption is available. Therefore, when the oxygen consumption is estimated for a participant not included in the training set, only the overall intercept term and the fixed regression structure are used.

We first fitted the model using all of the calculated features as explanatory variables. Then, we searched for the optimal model structure based on both the significance of the terms and a fit statistic of the model (-2 log likelihood ratio). The less significant terms (at a significance level of 0.1) were excluded from the model one by one. The fit statistic was monitored and the elimination of terms was stopped if it had a notable negative influence on the performance of the model.

We modeled the four activities separately. From the data set available from each activity, we excluded the measurements of one participant for testing the model and used the data from the other participants for feature selection and training the model. We chose a male participant (22 years, 86 kg, 1.81 m) for testing because the proportion of females in the data was low. In addition, this participant was chosen since his measurements in all the activities were available.



(a) Walking

(b) Running

(c) Nordic walking

(d) Bicycling

Fig 3. Measured (black) and estimated (grey) oxygen consumption. Reprinted with permission from Publication III ©2008 IEEE.

3.2.5 Results

Figures 3 a) - d) show the modeling results for the test data in the four activities. The black line in the figures represents the true oxygen consumption measured using indirect calorimetry, and the grey line is the estimated oxygen consumption given by the linear mixed model. The fact that the estimated curve starts later than the measured oxygen consumption curve results from the use of lagged effects in the model.

We can see in Figures 3 a) - d) that the model estimates the level of oxygen consumption very accurately, but the estimated oxygen consumption does not follow the measured oxygen consumption strictly. However, for all practical purposes, it is more important to find the correct level of oxygen consumption to be able to accurately estimate the total daily EE.

The actual EE can be calculated from the estimated oxygen consumption. The metabolic equivalent (MET) is a widely used unit for EE that represents the energy cost of physical activity as multiples of resting metabolic rate. One MET corresponds to an oxygen consumption of 3.5 ml/kg/min. Thus, the estimated oxygen consumption (VO_2 / kg) is converted to METs by dividing by 3.5. Table 5 shows a comparison of the average EE estimated by the linear mixed model with the average EE calculated from the true oxygen consumption measured using indirect calorimetry for each of the activities. In walking, running and Nordic walking, the model underestimated the average EE by 13, 2 and 9 percent, respectively, and in bicycling the average EE was overestimated by 7 percent.

Table 5. Measured and estimated energy expenditure in METs and the prediction error in four activities. Reprinted with permission from Publication III ©2008 IEEE.

	True	Predicted	Error
Walking	4.7	4.1	-13 %
Running	7.6	7.5	-2 %
Nordic walking	6.3	5.7	-9 %
Bicycling	6.8	7.3	+7 %

3.3 Cognitive load

Human attention is a finite resource, and the balance of our cognitive load or attention demands can easily fluctuate in situations of task interruption, divided attention and multitasking. For example, we can experience attention interference due to an unexpected interruption (e.g., a pedestrian crossing in front of a driver) or due to an expected interruption (e.g., a preset birthday reminder ringing on a cell phone during a conference presentation). Cognitive load might also increase when switching our attention between virtual spaces and physical spaces (e.g., using a navigation display while driving) or between two user interfaces (e.g., using a smart phone and a laptop together). In order to determine how to respond to the temporal and subtle changes of cognitive load, it is necessary to measure the cognitive load of individuals *in real-time* and *in situ*. With a real-time, objective measure, we can develop novel systems that can help users manage their cognitive effort and provide them with appropriate support.

3.3.1 Cognitive load modeling based on psycho-physiological measurements

There are reliable methods for assessing cognitive load, which are mostly based on task performance and subjective ratings (Cegarra & Chevalier 2008, Hart & Staveland 1988). However, due to their *post-hoc* (measured after an experience is complete) and static (measured at a single point in time) nature, these methods are inappropriate for measuring variations in cognitive load over a continuous time frame.

We chose the approach of modeling real-time cognitive load based on psycho-physiological measurements. Psycho-physiological measurements have been demonstrated in the literature as being useful for assessing cognitive load (*e.g.*, gaze information (Iqbal *et al.* 2005, Chen & Epps 2014, Piquado *et al.* 2010), heart rate (HR) (Koenig *et al.* 2011, Mehler *et al.* 2012), electroencephalography - the electrical activity of the brain (EEG) (Antonenko *et al.* 2010, Grimes *et al.* 2008, Knoll *et al.* 2011), electrocardiography - the electrical activity of the heart (ECG) (Koenig *et al.* 2011), galvanic skin response (GSR) (Koenig *et al.* 2011, Mehler *et al.* 2012), breathing rate (BR) (Koenig *et al.* 2011), and skin temperature (Koenig *et al.* 2011). As we were interested in identifying a generalized mechanism for assessing cognitive load, we wanted to stimulate that load using tasks that leverage basic cognitive processes instead of applied tasks, such as document editing (Iqbal *et al.* 2005), simulated public speaking (Fredericks *et al.* 2005), driving (Reimer *et al.* 2009, Mehler *et al.* 2009), learning (Antonenko *et al.* 2010), and aviation (Noel *et al.* 2005, Wilson 2002), used in earlier studies. In addition, earlier work in detecting cognitive load based on psycho-physiological measurements had not reached a granularity required for real-time assessment.

In Publication IV, we collected data from a range of psycho-physiological sensors from young adults and extracted a large number of features from the measured signals to explore which ones are useful for accessing cognitive load. We built individual cognitive load models for each of our participants. In Publication V, we further explored if we could perform the assessment in real-time.

Our cognitive ability changes with age (Salthouse 2010). As we age, deficits in cognitive abilities increase, and more support is needed (Craik & Rose 2012). However, the magnitude and speed of these changes also vary widely from person to person (Salthouse 2010). In Publication V, we also investigated if we can use the same sensors to measure cognitive load in older adults as for young adults.

3.3.2 Data set 1

For the first part of the study, we recruited twenty young adults with their age ranging from 19 – 34 (mean 25.15, sd 4.45), including 15 males and 5 females. They performed six elementary cognitive tasks (ECTs) that are basic tasks which require only a small number of mental processes, and which easily specify correct outcomes (Carroll 1993). We chose tasks that target ‘visual perception’ and ‘cognitive speed’ among the human cognitive abilities addressed in (Carroll 1993, McGrew 2009). These abilities engage spatial orientation or spatial attention (French 1951), which are highly leveraged in today’s world of location-based services, situations of divided attention, and applications where you may be attending to one activity and are either interrupted by incoming information or seeking information. The tasks we used were the *Gestalt completion test* (GC), the *Hidden pattern test* (HP), the *Finding A’s test* (FA), the *Number comparison test* (NC), the *Pursuit test* (PT), and the *Scattered X’s test* (SX) (see Figure 4). The participants used a mouse and a keyboard to answer the screen-based ECT questions.

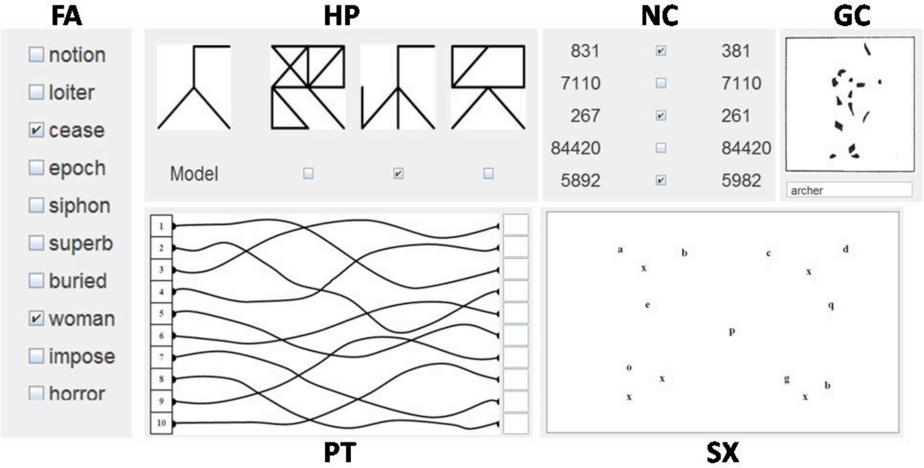


Fig 4. Six elementary cognitive tasks (ECTs). Reprinted with permission from Publication IV ©2010 ACM.

For each ECT, two sets of questions were shown to the participant in a random order. One of the sets contained questions of a lower difficulty level (inducing a lesser degree of cognitive load), while the other was comprised of more difficult questions (inducing a greater degree of cognitive load). The participants were given up to 3 minutes to answer each set of questions.

While completing the tasks, the participants wore three sensor devices (the NeuroSky Mindset wireless EEG headset, the Polar RS800CX heart rate monitor, and the SenseWear Pro3 armband) measuring EEG, ECG, cardiac inter-beat (R-R) intervals, GSR, and heat flux (rate of heat transfer on the skin). In addition, the participants executed all the tasks in front of a contactless eye tracker (SmartEye 5.5.2), comprised of two cameras, which measured pupil diameter. The experiment setup and sensor devices are demonstrated in Figure 5. Figure 6 shows examples of the measured signals.

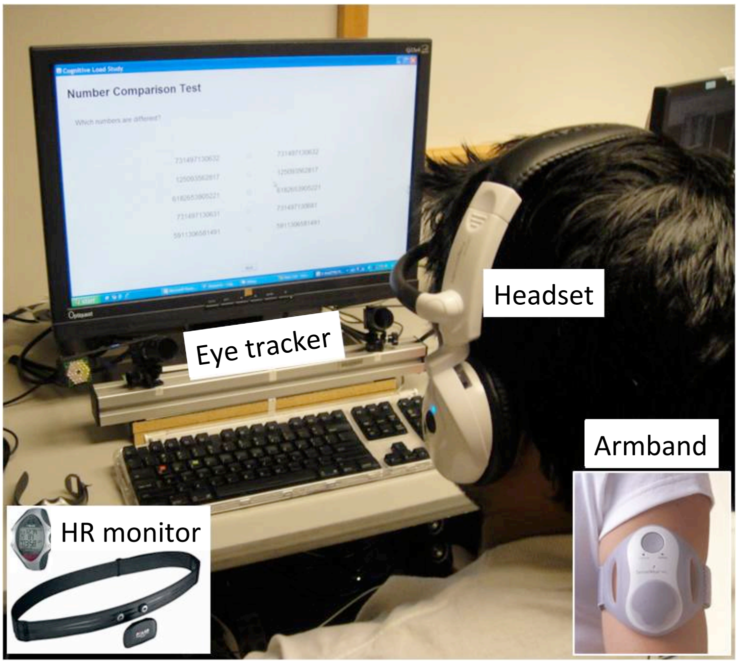


Fig 5. Experiment setup and sensor devices. Reprinted with permission from Publication IV ©2010 ACM.

Before analysis, we preprocessed the heart rate R-R data by removing outliers falling outside the range of 35 – 155 bpm (387 – 1714 ms). An increasing trend was removed from the GSR signals and the lowest and highest 0.1 percent of values from each participant were excluded as outliers.

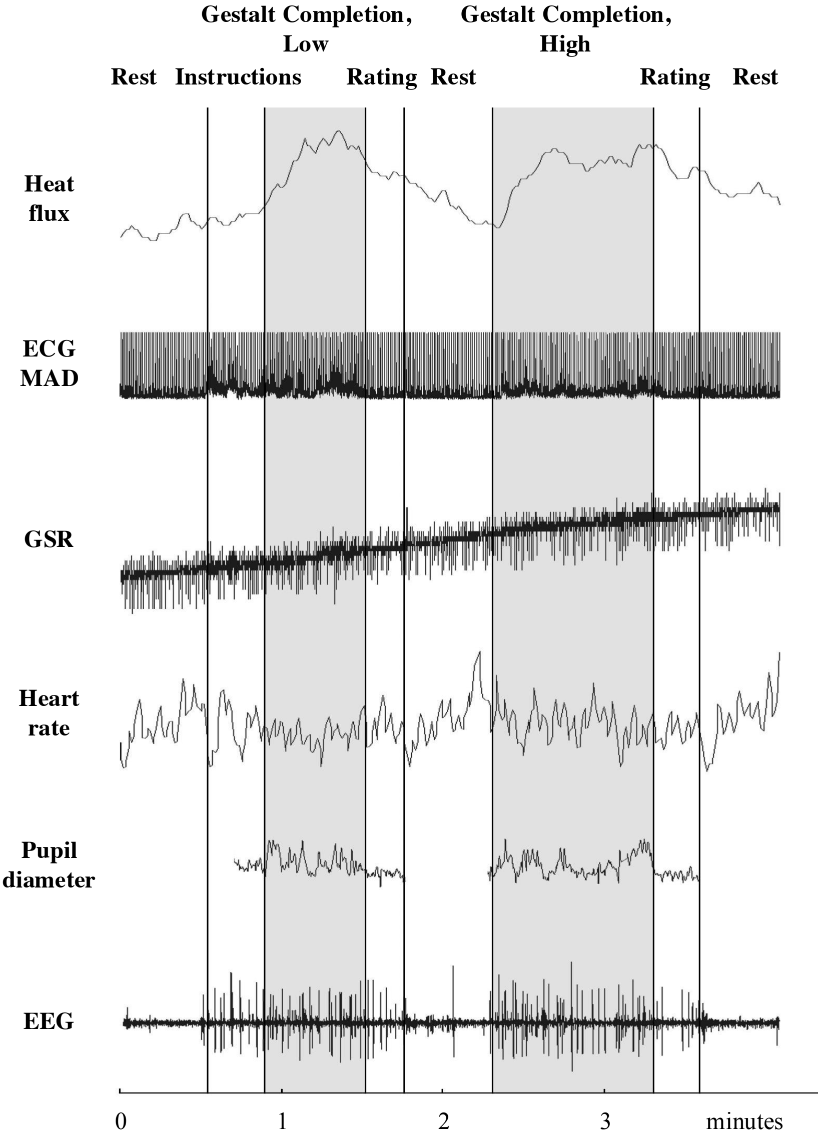


Fig 6. Example psycho-physiological signals collected during the Gestalt Completion test (low and high difficulty). Reprinted with permission from Publication IV ©2010 ACM.

3.3.3 Features

We modeled the level of cognitive load (low vs. high) using features derived from non-overlapping segments of psycho-physiological sensor data corresponding to the different questions in the ECT tests. The average length of the segments was 23.7 seconds although the length varied from a few seconds to several minutes. The altogether 51 statistical features included the mean, variance and median of pupil diameter, GSR, heat flux, ECG median of absolute deviation (MAD – a measure of variability of ECG), 8 EEG power values and two mental state outputs (provided by the EEG headset) as well as the average spectral power of the raw EEG signal on five bands (delta $0-4$ Hz, theta $4-7$ Hz, alpha $8-12$ Hz beta $12-30$ Hz and gamma over 30 Hz). Two heart rate variability (HRV) features, standard deviation of R-R intervals (SDNN) and the root mean square of the successive differences of R-R intervals (RMSSD), as well as the mean and variance of HR were derived from the HR data.

3.3.4 Evaluation of individual features

We then evaluated the performance of each of these features in assessing cognitive load. Because of individual differences in the levels of psycho-physiological responses to cognitive load, we modeled each participant individually. For each question type, the data from the separate questions were classified into one of two classes representing the two difficulty levels. Classification was performed based on one feature alone, using a Naïve Bayes classifier (NB). We used a leave-one-out validation approach between the questions in each question type to calculate the average classification accuracy for the question type. Because the difficulty levels in two different question types were unlikely to correspond to each other (*e.g.*, high difficulty questions for the Finding A's and the Pursuit tasks would not necessarily induce the same amount of cognitive load in a participant), only data from the same question type was used in the classification. The overall classification accuracy of the feature was then calculated as the average accuracy over the 6 question types. This was performed for each participant and each feature in turn.

Table 6 shows the best feature and the corresponding classification accuracy for each participant. The results show that for each participant, a feature that discriminates the two classes with a high accuracy was found. Most of the best features were calculated from either the heat flux measurement or the ECG signal.

Table 6. The best feature for each participant and the corresponding classification accuracy. Revised from Publication IV ©2010 ACM.

Participant	Average accuracy (%)	Best feature
1	82.5	Heat flux - median
2	86.7	ECG MAD - median
3	86.7	Heat flux - mean
4	74.0	ECG MAD - median
5	81.7	EEG power low beta - median
6	76.3	EEG attention - mean
7	83.3	Heat flux - median
8	80.4	Heat flux - median
9	86.3	EEG power high beta - median
10	87.0	Heat flux - median
11	92.5	ECG MAD - median
12	75.5	GSR - variance
13	78.3	ECG MAD - mean
14	80.8	Pupil diameter - median
15	82.5	ECG MAD - median
16	81.3	ECG MAD - median
17	88.3	ECG MAD - variance
18	89.2	Heat flux - mean
19	94.0	Heat flux - mean
20	76.3	EEG power theta - variance

Table 7 presents the average classification results over all 20 participants using models created with the best feature from each sensor stream. Here again, the features that perform the best are based on either the heat flux measurement or the ECG signal. The classification performance of features calculated from the other measurements is clearly inferior.

The two best features (median of heat flux and median of ECG MAD) were then used together to classify the levels of cognitive load. The average classification accuracy across participants was 81.1%, which is higher than the accuracy of using any single feature alone.

Table 7. Average classification results of the best features from each sensor stream over 20 participants. Revised from Publication IV ©2010 ACM.

Sensor	Feature	Average accuracy (%)
Heat flux	median	76.1
ECG MAD	median	71.4
EEG attention	median	60.2
HR	mean	58.7
Pupil diameter	median	57.4
GSR	variance	53.7

3.3.5 Data set 2

In continuing the study (Publication V), we wanted to explore if we could perform the assessment in real-time and also, if the same psycho-physiological measures would be indicative of cognitive load for older adults as for young adults. This is particularly challenging, given the changes in psycho-physiological responses that occur as a part of aging (Anderson & McNeilly 1991). We recruited 30 participants split across two age groups: 13 young with their age ranging from 18 – 30 (mean 22.9, sd 3.9), including 6 males and 7 females, and 17 older participants with age ranging from 65 – 88 (mean 74.3, sd 5.7) including 6 males and 11 females. From the six ECTs used in the earlier stage of the study, we this time presented two to the participants: the Pursuit test (PT) and the Scattered X's test (SX).

Again, we prepared two sets of questions for each question type with different difficulty levels. The questions were presented to the participants in three separate blocks with each block containing one question set of each difficulty level for each of the two ECTs. The duration of the question sets in the first block was 4 minutes, 3 minutes in the second, and 2 minutes in the last block.

We measured participants' psycho-physiological responses with four sensor devices, two of which were the same as in the first study and two were newly introduced. The SenseWear Pro3 armband was used to measure heat flux and the NeuroSky Mindset was used to record EEG activity. The two new sensor devices were a wireless ECG monitor (Bioharness BT), that also records heart rate (HR) and breathing rate (BR), and a GSR finger sensor (LightStone). We preprocessed the raw GSR and heat flux signals by convoluting with a Bartlett window (Oppenheim & Schafer 2010) to smoothen, and by differencing to remove trend.

We encountered some challenges with noisy or missing sensor data, especially with the EEG and ECG sensors. The poor quality of the EEG signal may be caused by poor contact of the sensor/ground/reference electrodes to a participant's skin, motion of the participant, environmental electrostatic noise, or non-EEG biometric noise (*i.e.*, EMG, ECG, EOG, and others) (NeuroSky 2009). Sources of recording noise in the ECG can include artifacts caused by movement of the electrode away from the contact area on the skin, or EMG noise due to muscle contractions under the sensor surface (Friesen *et al.* 1990). Particularly, ECG measurements from the older participants had irregularities that may have influenced the extraction of R-R intervals. Participants who had missing or noisy ECG and EEG data were excluded from the analysis (2 young subjects and 5 older subjects, highlighted in grey in Table 8). In addition, the heat flux signal was missing from one participant due to a device malfunction and the GSR measurement had to be dropped from 5 participants because of incomplete readings. This left us with 11 young and 12 older participants with whom we continued the analysis.

3.3.6 Features

We generated two separate feature sets from the raw measurement signals to study cognitive load assessment at two different granularities. In the first feature set, raw measurement signals were represented by statistical features calculated on 60-second sliding windows, and in the second set on 10-second windows both with a step of one second.

In total, we extracted 128 features from the signals. The mean, median, variance, standard deviation, 10th, 25th, 75th and 90th percentile, interquartile range (IQR), RMSSD, mean of the absolute values of the first (MAFD) and second (MASD) differences, mean crossing rate and the difference of the last second mean and the first second mean of the window (end-start difference) were calculated from the heat flux, GSR and the R-R interval signals. Further, correlation and SDNN, relative occurrence of successive differences exceeding 20ms (pNN20) and 50ms (pNN50), and mean peak amplitude were calculated from the R-R data. The count, maximum amplitude, mean amplitude, mean duration and area under skin conductance response (SCR) occurrences were extracted from the GSR measurement. The heart rate, breathing rate and breathing wave amplitude were described by seven features: minimum, maximum, mean, median, variance, standard deviation and end-start difference. We calculated spectral power of the raw EEG data on five bands: delta, theta, alpha, beta and gamma.

We normalized all the features to equalize their importance, and scaled the feature data from each participant in each question type linearly so that the 5th and 95th percentiles of each of the features met the range $[0,1]$.

3.3.7 Modeling and feature selection

We used quadratic discriminant analysis (QDA) to classify the feature values calculated on the sliding windows. We adopted a block study design similar to the cross-validation scheme suggested by Grimes *et al.* (2008) to avoid temporal dependence of the data segments and distortion of the results. We used the three data sets corresponding to the three blocks of the study design to train the models, select subsets of the original set of 128 features that would give the highest accuracy in measuring cognitive load, and to simulate a real-time system to provide an estimate of how the model would perform on a previously unseen set of data, respectively. For feature subset selection, we used the simple algorithm of selecting the three best individual features. We tested also more sophisticated methods, but because of overfitting, this simple method proved the most efficient for the task.

We trained a model with individual feature selection for each participant's data for each ECT task to distinguish the two difficulty levels. Cognitive load assessment accuracies for the young and older participants are presented in Tables 8 and 9, respectively. The assessment was performed at a one-second frequency. Very high accuracies are achieved for many of the young participants, but with high variation. The rationale for the models not working for all participants might include noise in the measurement data or individual differences in how the changes in cognitive load manifest themselves in psycho-physiological signals. Participants might also have experienced the two levels of task difficulty differently, even though we did not find clear evidence of this in subjective ratings and task performance recorded during the data collection. Also our decision to fix the number of features at three might explain the inferior accuracies for some of the participants. This parameter value resulted in the best results for most of the participants, but some of them would have benefitted from a higher number of features in the model.

Table 8. Sensor signal quality and cognitive load assessment accuracies in the two ECT's for young participants. Revised from Publication V ©2014 IEEE.

					PT		**SX**	
#	Heat flux	GSR	EEG	ECG BR, HR	10s	60s	10s	60s
1	●	●	●	○	30%	53%	44%	66%
2			●	●	71%	**100**%	**88**%	39%
3	●		○					
4	●	●		●	**77**%	**100**%	70%	50%
5	●	●	●	●	**94**%	**86**%	**100**%	**100**%
6	●	●		●	50%	50%	62%	**89**%
7	●	●	●	○	**76**%	**100**%	**76**%	69%
8	●	●	●	○	49%	54%	63%	31%
9	●	●	●	○	**76**%	53%	63%	**85**%
10	●	●	●	○	**79**%	74%	45%	**86**%
11	●	●		●	74%	**95**%	**98**%	**100**%
12	●	●	●	○	52%	**100**%	**92**%	**100**%
13	●	●		○				
				Avg	66%	**79**%	73%	74%

● - Good ○ - Poor - Missing

The accuracies for the longer windows are generally better than for the shorter windows in the PT task, but in the SX task the assessment performance is equal at both granularities. Even the 10-second assessment is able to differentiate the two levels of cognitive load at a very high accuracy for some of the participants. The results are particularly good for young participants 5, 11 and 12. On the other hand, our models did not work for young participants 1, 6, and 8. However, participant 6 was missing EEG data, and participants 1 and 8 had poor ECG, HR and BR data. The results for the older participants are very similar, and the 60-second granularity results are comparable to those of the young adults. However, for the older adults, the difference in accuracy between the two granularity levels is greater than for the younger participants. This might be a consequence of the psycho-physiological changes related to aging (Anderson & McNeilly 1991). Our models did not work well for older participants 2, 5, 10 and 13. We suspect that the reason for this might be that some older adults do not express cognitive function with a high enough signal through psycho-physiological responses.

Table 9. Sensor signal quality and cognitive load assessment accuracies in the two ECT's for older participants. Revised from Publication V ©2014 IEEE.

					PT		SX	
#	Heat flux	GSR	EEG	ECG BR, HR	10s	60s	10s	60s
1	●	●		○				
2	●	●	●	●	57%	**91**%	63%	66%
3	●	●		●	62%	**79**%	63%	**79**%
4	●	●		●	53%	**81**%	**77**%	**77**%
5	●	●	●	●	62%	**100**%	62%	52%
6	●	●	●	●	**88**%	**98**%	55%	50%
7	●	●		○				
8	●	●	●	●	49%	**85**%	**77**%	**100**%
9	●	●	●	●	69%	**95**%	**90**%	58%
10	●	●	●	○	54%	50%	50%	24%
11	●	●		●	72%	**99**%	61%	**99**%
12	●		●	○	51%	53%	**80**%	**100**%
13	●	●	●	●	72%	**100**%	50%	50%
14	●	●		○				
15	●			○				
16	●		●	●	**80**%	**96**%	50%	28%
17	●	●		○				
				Avg	64%	**86**%	65%	65%

● - Good ○ - Poor - Missing

We then analyzed the features selected for each of the models at the 10-second granularity. The relative count of times a feature was selected into the set of three best features, normalized by the number of participants in the age group, from which that sensor stream was available, is shown in Figure 7. For this, we only considered the participants and tasks that had an accuracy over 75%. The most often selected features for both the young and the older participants originate from the EEG and BR signals. The EEG signal was more important and the BR measurement less important for the older than for the young participants. R-R signals are also fairly well represented among the most common features for both age groups, whereas GSR features were seldom used for the young participants, and never for the older participants.

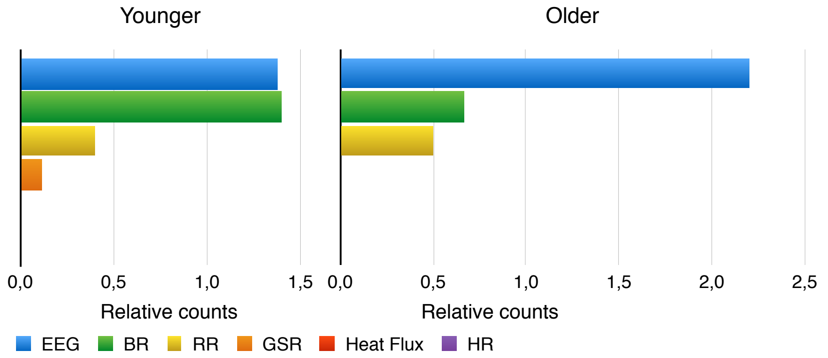


Fig 7. Relative count of the times a feature from each sensor stream was selected into the models (both tasks, 10-s). Reprinted with permission from Publication V ©2014 IEEE.

Even though our models did not work for all of the participants, our results showed that we were able to model cognitive load in real-time for both young and older adults, and that we can use the same sensors for both age groups.

3.4 Summary

This chapter presented the results of our three case studies. In the first study, we classified RSW processes to find process parameters and quality control methods to set up a new welding process. We compared eight different sets of features calculated from current and voltage signals recorded during welding with different compositions, dimensionality reduction levels and with, or without, standardizing the data. Five different classifiers were tested. The best results were obtained with the *3*NN classifier and a non-standardized feature set consisting of 10 means calculated at intervals of equal length on the signals. We further improved the results by applying feature selection algorithms to the original set of 54 features. Five different feature selection methods were tested and the best results with a very high classification accuracy were obtained with a feature set consisting of 11 features selected by the SFFS algorithm.

The second study focused on modeling physical activity in four exercises based on body-acceleration. To account for the large deviation in exercise EE between participants, we used the linear mixed model that models both between-subject and within-subject variation. The feature set we used for the model consisted of features based on the variance of the acceleration signal, as well as features based on the participants' height.

Data from one of the participants was kept for testing the model performance while the linear mixed model was trained using data from the remaining participants. For each of the four activities, a subset of the features was selected based on both the significance of the terms in the model and a fit statistic. Our models were able to accurately estimate the EE.

In the third case study, our aim was to assess human cognitive load based on psycho-physiological measurements. In the first stage of the study, we collected a range of psycho-physiological signals from young adults while they were solving computer-based ECT tasks with two levels of induced cognitive load and calculated altogether 51 statistical features from these signals. We then evaluated the performance of each of the features individually using the NB classifier. We used leave-one-out validation between the questions in each task type to calculate the average classification accuracy for the task, and then averaged over the six task types to obtain an average classification accuracy for each feature. Our results showed that for each of our participants, we were able to find a feature that discriminates the two classes with a high accuracy. We also found that combining the two individually best features, originating from the heat flux and ECG MAD signals, resulted in a very high classification accuracy across all participants.

In the second stage of the study, we explored if we could perform the assessment in real-time and if the same psycho-physiological measures would be indicative of cognitive load for older adults. We collected a data set with psycho-physiological measurements from both young and older adults and extracted 128 features from the signals at two different time granularities, 10 seconds and 60 seconds. We used a block cross-validation scheme for training, feature selection and testing our models built with the QDA classifier on a feature set selected by the *3*Best features selection algorithm. We showed that we are able to model cognitive load in real-time for both young and older adults, and that we can use the same sensors for both age groups.

4 Discussion

In this thesis, we have addressed different aspects of the machine learning process, including data preparation, selection of the model type, feature selection and validation with a special emphasis on working with time series data. We have also presented our work and results in three different time series machine learning application areas: resistance spot welding process identification, exercise energy expenditure estimation, and cognitive load modeling. In this chapter we will first summarize our findings from each application area, and then discuss the contributions of our work as a whole from both a theoretical and a practical point of view.

4.1 Case studies

We will now discuss our choice of methods and our findings in each individual case study, and also consider the limitations of our work.

4.1.1 Resistance spot welding

In our first case study, resistance spot welding (Publications I and II), we found that the non-parametric methods we tested worked better than the parametric models for classifying the different welding processes. We can assume that this might be explained by the diversity of the data. The non-parametric methods performed better because they do not make assumptions about the data distribution, whereas the parametric models we used assume the data to originate from normal distributions. We also noticed that even though standardization is usually a recommended step of the data preparation process (Gelman & Pardoe 2007), it weakened the classification accuracies in this case. This implies that the features calculated from the current signal, which originally had the wider measurement range, were more important for the classification than the features calculated from the voltage signals. This reminds us of the advice given by Hand *et al.* (2001) that the algorithms selected for data analysis should always be tailored for the particular problem at hand, rather than blindly following a commonly used procedure. More importantly, understanding the data lays the foundation for solving any machine learning problem.

We also searched for the optimal parameter values for the classifiers using 10-fold cross-validation in the training set, and applied several feature selection methods to improve the classification accuracy. We found the forward methods, SFFS and SFS, to result in the best feature sets. However, our research unfortunately fell into the common pitfall described in Section 2.4 (Kohavi & Sommerfield 1995, Ambroise & McLachlan 2002, Reunanen 2003) as we presented the validation set accuracies of our candidate models as final performance estimates of the models. To obtain unbiased estimates, we should have used a test set separate from the model selection process. Hence, the feature subset found with the computationally more thorough methods might actually not be better for distinguishing the different welding processes than the more simple *n*Best method that also produced very good results. Also the classification accuracies we presented might be somewhat higher than what we would have obtained with a fresh data set. However, according to Kohavi & Sommerfield (1995) and Cawley & Talbot (2010), overfitting is mainly a problem when the data set available is small. Hence, our relatively large data set and the fact that the accuracies for all of our models, with or without, feature selection were similar, ensure the validity of our results in general.

4.1.2 Exercise energy expenditure

In the second case study of this work, exercise energy expenditure modeling (Publication III), we chose the linear mixed model structure over the general linear model to account for the temporal dependency of the time series measurement data. Particularly, the repeated measurements from each participant are correlated with each other, whereas the general linear model assumes observations to be independent. We found that with this model structure, we were able to account for individual differences in the oxygen consumption and produce accurate estimates of energy expenditure in four different activities.

We should mention as a shortcoming of our study that we only tested our models on the data from one of the participants. However, these preliminary results were enough to demonstrate the feasibility of the new approach to exercise energy expenditure modeling.

4.1.3 Cognitive load

In the third case study (Publications IV and V), we searched for a method to assess cognitive load for both young and older adults in real-time. We derived a large number of features from psycho-physiological measurement signals, and in the first stage of the study evaluated their individual value in modeling cognitive load. We presented cross-validation results where we used non-overlapping segments of the measurement data. However, as in the spot welding study, we only presented the leave-one-out validation accuracies. In addition, since the data segments corresponding to the two difficulty levels of the tasks were collected adjacently, and only one data set was collected, these results are likely to be optimistically biased. In fact, since only one data set was collected, it would not have helped to put aside a part of the data for validation; the temporal dependency of the time series data would still have affected the results.

We corrected this mistake in the second stage of the study where we used a block study design adapted from Grimes *et al.* (2008). We used three blocks of data collected separately to train our models, select the best features and evaluate our results. Although this time our data collection suffered from quite severe problems with the measurement devices and data quality, these results showed that we were able to assess the cognitive load of both young and older adults on a fine time granularity.

Since the data sets we used for this study were relatively small, we chose to use simple parametric classifiers. Also for the feature selection, we chose the simple *3*Best method since we found the use of more complex methods to result in overfitting of the models.

The measurement signals that we found to be the most valuable for assessing cognitive load in these two publications were different as the heat flux and ECG appeared most informative of cognitive load in Publication IV, whereas EEG, BR and RR were selected as the best signals in Publication V. However, the RR, in fact, is derived from the ECG signal. The changes in cognitive load probably did not manifest themselves in the heat flux signal fast enough for the latter part of the study where the assessment was carried out at a more fine-grained temporal granularity.

4.2 Theoretical implications

The objective of this thesis was to discuss aspects of the model selection process that should be considered when solving a machine learning problem in general, and when working with time series data in particular. We will now discuss the three research questions we set in Chapter 1 based on our case studies, and compare our findings with the literature reviewed in Chapter 2.

1. **How should the data characteristics and the amount of data available guide the selection of model type and the model selection strategy?**

Based on the characteristics of the data at hand, we can draw some general guidelines for selecting the approach to use at the different steps of the machine learning process. From our case study in resistance spot welding, we learned that if the data set available is relatively large and we do not know the distribution of the data (Publications I and II), a nonparametric model might work better than a parametric one. On the other hand, if the data set is limited, a parametric model might still work well, even though the data might not follow the distribution assumed by the model (Publications IV and V). These findings are also well-known in the literature (Hand *et al.* 2001, Duda *et al.* 2001, Bishop 2006).

Over-fitting in model selection has been a widely discussed topic during the past two decades (Wolpert 1992, Schaffer 1993, Kohavi & Sommerfield 1995, Reunanen 2003, 2004, Raudys 2006, Cawley & Talbot 2010). In Publications IV and V we concluded, in agreement with Raudys (2006), that simple classification algorithms worked better with small data sets, whereas the use of more complex models would have resulted in overfitting the data.

In addition, the selection of a feature selection algorithm depends on the amount of data available. To avoid over-fitting, and to ensure the generalizability of results, Reunanen (2003) recommends using more simple feature selection strategies when little data is available. In line with this, we found the simple *3*Best method to perform best for our cognitive load modeling study (Publication V).

2. **How should the data set available be used for model training, selection and validation to optimize the model generalizability and performance on new data?**

The importance of using a separate test set in evaluating the selected model has been recurrently brought up in the literature (Kohavi & Sommerfield 1995, Reunanen 2003, Fiebrink & Fujinaga 2006, Smialowski *et al.* 2010). Nevertheless, new studies are repeatedly published where the training set or cross-validation estimates are presented as the final performance estimates of the model (Fiebrink & Fujinaga 2006, Cawley & Talbot 2010). Unfortunately, our Publications II and IV also fell into this pitfall, but the mistake was corrected in Publication V. Even though this finding is not new as such, the importance of always using a data set previously unseen by the model for performance evaluation in any machine learning application cannot be emphasized too much.

3. **What are the special considerations to take into account when applying machine learning algorithms to time series data?**

The most important thing to keep in mind when working with time series data is the ordering of the observations (Pyle 1999). Features can be calculated on subsequent segments of the original time series, and modeling algorithms can be selected that take the correlation of the data into account. In Publication III, we chose the linear mixed model over the general linear regression model to account for the temporal dependency of the repeated oxygen consumption measurements from our participants.

More importantly, however, the temporal dependency of time series data needs to be remembered when dividing a data set for training, model selection and evaluating the performance of the final model. While the dependency of the adjacent observations in a time series is widely noticed, and this characteristic is taken into account when preparing data sets in statistics (Hart & Wehrly 1986, Burman *et al.* 1994, Arlot & Celisse 2010), it seems to be less acknowledged in the field of machine learning (Grimes *et al.* 2008). Our work highlights the importance of this in Publication V, where we used a block design in data collection where separate contiguous sequences of observations were used for training, testing and validating our models.

Overall, our work shows that for the most part, the commonly used machine learning methods can also be applied in problems involving time series data. However, our results highlight the importance of considering the temporal dependency of the data, both when selecting a modeling algorithm to be used, and when partitioning the data set available for training, model selection and estimation of the model performance.

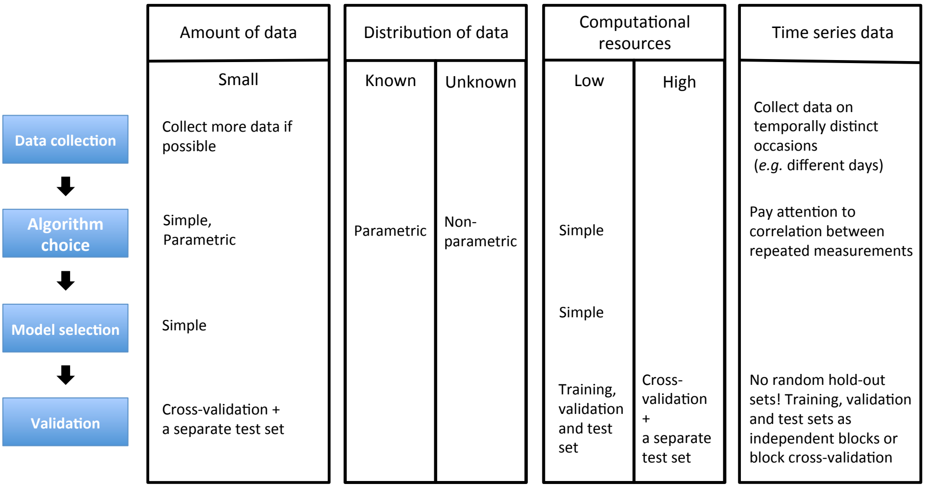
4.3 Practical implications

In this thesis, we have presented our work in three different machine learning application areas where we solved modeling problems including time series data. In resistance spot welding, we identified new welding processes with a high accuracy to match them with process parameters and quality control methods from previous similar processes. This information can then be used to speed up the set up of a new welding process, thereby obtaining both material and financial savings. In exercise energy expenditure estimation, we introduced the use of the linear mixed model to model energy expenditure, based on measurements from a wrist-worn accelerometer. The use of a wrist-worn sensor eases the use of the monitoring system compared with the traditional chest bands used for exercise energy expenditure estimation. In cognitive load modeling, we assessed the cognitive load of both young and older adults in real-time, based on psycho-physiological measurements. Our study was one of the first ones to reach a fine temporal granularity, and the first one to compare such a variety of off-the-shelf measurement devices for cognitive load assessment.

In addition to our contributions in each of these application areas, the considerations we brought up in this thesis might help a new practitioner in the field of machine learning to adopt a sound methodology and not fall into the pitfalls we, as well as many more senior researchers, have fallen. In particular, we hope that our work will help the machine learning community to learn to apply the methods more cautiously to prevent the negative effects of model over-fitting and feature over-selection, and to better acknowledge the temporal dependency of time series data, especially when partitioning data for training, selecting and evaluating models, to ensure unbiased and comparable results.

As our final contribution, we would like to offer Figure 8, where we summarize the findings of this thesis. This figure can be used as a checklist for anyone starting to tackle a new machine learning problem. It covers the four steps of the machine learning process that we discussed in this thesis: *data collection*, *algorithm choice*, *model selection* and *validation*. In this figure, we highlight four dimensions of a machine learning problem to consider when deciding on an approach to take: the *amount of data*, the *distribution of the data*, *computational resources* and whether the problem involves *time series data*.

More precisely, we advise consideration of whether the distribution of the data is known or unknown, which guides the choice of a parametric or non-parametric method, and whether the computational resources available are low or high, which determines if a simple or a more complex algorithm should be used. For the amount of data we only consider the case where the size of the data set is limited, because in the case of a large data set the choice of methods is predominantly only limited by the computational resources. In the case of a small data set, we recommend, if possible, to collect more data, as it is probable that a greater improvement in the performance of any algorithm would be obtained by further data collection than by selection of a more optimal algorithm (Cawley & Talbot 2010). Also the quality of the data might be more crucial for the success of the modeling effort than the selection of a single algorithm. Hence, the data collection step should be especially invested in, particularly when collecting real-world time series data, to minimize noise, missing measurements and other problems with the data quality. If, however, the data set available is small



	Amount of data	Distribution of data		Computational resources		Time series data
	Small	Known	Unknown	Low	High	
Data collection	Collect more data if possible					Collect data on temporally distinct occasions (*e.g.* different days)
Algorithm choice	Simple, Parametric	Parametric	Non-parametric	Simple		Pay attention to correlation between repeated measurements
Model selection	Simple			Simple		
Validation	Cross-validation + a separate test set			Training, validation and test set	Cross-validation + a separate test set	No random hold-out sets! Training, validation and test sets as independent blocks or block cross-validation

Fig 8. Four dimensions of a machine learning problem to consider when choosing an approach to take for the four steps of the machine learning process discussed in this thesis.

and additional data collection is not feasible, it is recommended to choose simple algorithms to avoid over-fitting in model training or selection. Parametric methods often work better than non-parametric ones for small data sets. Cross-validation is a good procedure to make the most of a small data set. However, if any model selection with several candidate models is involved, testing with a separate data set should not be forgotten. The special nature of time series data should also be remembered when choosing modeling algorithms to use, and when dividing the available data for training, model selection and testing the performance of the models.

Based on our experience with real-world applications, we can also reflect on the requirements set by the nature of the particular problem. One of our application areas, resistance spot welding, took place in an industrial setting, whereas the other two applications, exercise energy expenditure estimation and cognitive load modeling, involved the human aspect. In an industrial application, where the data is collected from machinery in operation, data collection is often relatively cheap and reliable. In this case, even a small improvement in model performance, achieved either through additional data collection or fine-tuning the model, can bring large financial savings through improved manufacturing efficiency in the long run. In applications involving data collection from human participants, however, the data collection is generally more laborious and costly. Also the quality of data is often more variable due to sensor placement and reliability issues as well as personal differences between the participants.

In addition to the computational resources available during the model training stage, also the end-use circumstances of the model need to be considered when choosing a model to use. In two of our applications, resistance spot welding and cognitive load modeling, the final model was run on a desktop computer where computational processing power can be increased when necessary. The exercise energy expenditure models, on the other hand, were run on a wearable sports watch where the storage and computational resources limit the complexity of the model.

The requirements for model accuracy also vary according to the application area. In resistance spot welding, for example, malfunctioning of a model can result in rejection of a whole production batch, hence causing great financial and material losses. In a safety critical application relying on cognitive load modeling, the consequences of a single wrong decision made by a model can be very severe. In exercise energy expenditure modeling, on the other hand, the correctness of an energy expenditure estimate at a single moment in time is rather insignificant, while finding the correct level of energy expenditure during an exercise, and thereby being able to calculate the average energy expenditure for the activity or the whole day, brings more value for the user.

In summary, in addition to the general guidelines that can be given based on the data characteristics, amount of data, computational resources and potential time series nature of the problem, also the requirements of the environment in which the final model will be used need to be considered when selecting models for a specific application. Hence, based on our experience with practical machine learning applications, we emphasize the importance of Hand *et al.*'s (2001) statement that the methods used at the different steps of the machine learning process should always be selected for the particular problem at hand, rather than adopting a commonly used combination.

5 Conclusions

Model selection is a necessary step for any practical machine learning task. Since it is impossible to know the true model behind any real-world process, the goal of model selection is to find the best approximation among a set of candidate models. In practice, this means finding a model that best describes the real-world process to be modeled or best predicts the future outcomes of the process. In the case of predictive modeling, a crucial property of the model is its generalizability to new data. Therefore, testing with a separate data set should always be included as a part of the model selection procedure.

However, sound methodology is often forgotten and research results are repeatedly published where model accuracy estimates are based on the data set used for model selection. In this thesis, we brought forth problems relating model over-fitting and over-selection caused by careless or uninformed application of model selection methods. We discussed model selection in time series machine learning applications, and presented our results in three different application areas: resistance spot welding process identification, exercise energy expenditure estimation and human cognitive load modeling.

In each of the individual application areas, we developed models to solve a real-life machine learning problem. We covered the whole machine learning process including data collection and preparation, selection of the model type, feature selection and validation. Overcoming challenges typical to real-world modeling tasks, such as sensor reliability issues, individual differences between people, and requirements for the ease of use of monitoring systems and the temporal granularity of detection, we obtained accurate models that can be implemented in end-user applications and products.

Based on our findings in these studies we drew general guidelines on the points to consider when starting to solve a new machine learning problem and also discussed how the nature of the problem at hand affects the choice of algorithms to use. Throughout the thesis, we paid additional attention on the special nature of time series data and the restrictions and requirements it sets on the methods selected.

For future work, it would be interesting to study how our results on resistance spot welding process identification would change if we ran the study again following the recommendations for dividing the data set for model training, selection and evaluation discussed in this thesis. However, due to the project nature of our work, going back to an old study is not seen to be worthwhile. Instead, we will keep these lessons in mind when solving new machine learning problems in the future. It is our hope that our work will also serve other researchers in the field of machine learning as a reminder of cautious application of model selection algorithms. From the variety of application areas covered, we can see that the considerations brought forth in this thesis apply to machine learning modeling tasks in general.

References

Agrawal R, Faloutsos C & Swami A (1993) Efficient similarity search in sequence databases. Proc. 4th International Conference on Foundations of Data Organization and Algorithms, 69–84.

Albinali F, Intille S, Haskell W & Rosenberger M (2010) Using wearable activity type detection to improve physical activity energy expenditure estimation. Proc. 12th ACM International Conference on Ubiquitous Computing, 311–320.

Ambroise C & McLachlan GJ (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. Proceedings of the National Academy of Sciences 99(10): 6562–6566.

Anderson NB & McNeilly M (1991) Age, gender, and ethnicity as variables in psychophysiological assessment: Sociodemographics in context. Psychological Assessment: A Journal of Consulting and Clinical Psychology 3(3): 376.

Anderson T (2001) Radiographic and ultrasonic weld inspection: Establishing weld integrity without destroying the component. Practical Welding Today, December 13 .

Antonenko P, Paas F, Grabner R & Gog T (2010) Using electroencephalography to measure cognitive load. Educational Psychology Review 22(4): 425–438.

Arlot S & Celisse A (2010) A survey of cross-validation procedures for model selection. Statistics surveys 4: 40–79.

Bishop CM (2006) Pattern recognition and machine learning. Springer-Verlag, Secaucus, NJ.

Blair SN (2009) Physical inactivity: The biggest public health problem of the 21st century. British Journal of Sports Medicine 43(1): 1–2.

Borovkov AA (1999) Mathematical statistics. Gordon and Breach Science Publishers, Amsterdam.

Bouten CV, Sauren AA, Verduin M & Janssen JD (1997) Effects of placement and orientation of body-fixed accelerometers on the assessment of energy expenditure during walking. Medical and Biological Engineering and Computing 35(1): 50–56.

Bouten CVC, Westerterp KR, Verduin M & Janssen JD (1994) Assessment of energy expenditure for physical activity using a triaxial accelerometer. Medicine and Science in Sports and Exercise 26: 1516–1523.

Burman P, Chow E & Nolan D (1994) A cross-validatory method for dependent data. Biometrika 81(2): 351–358.

Carroll JB (1993) Human Cognitive Abilities: A Survey of Factor-Analytic Studies. Cambridge University Press, Cambridge, MA.

Cawley GC & Talbot NL (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research 11: 2079–2107.

Cegarra J & Chevalier A (2008) The use of Tholos software for combining measures of mental workload: Toward theoretical and methodological improvements. Behavior Research Methods 40(4): 988–1000.

Chatfield C (2003) The analysis of time series: An introduction. Chapman & Hall/CRC, Boca Raton, FL.

Chen KY & Bassett J David R (2005) The technology of accelerometry-based activity monitors: Current and future. Medicine and Science in Sports and Exercise 37(11 Suppl): S490–S500.

Chen S & Epps J (2014) Using task-induced pupil diameter and blink rate to infer cognitive load. Human–Computer Interaction 29(4): 390–413.

Cho Y & Rhee S (2002) Primary circuit dynamic resistance monitoring and its application to quality estimation during resistance spot welding. Welding Researcher 81(6): 104–111.

Chu CKK & Marron JS (1991) Comparison of two bandwidth selectors with dependent errors. The Annals of Statistics 19(4): 1906–1918.

Cover T & Hart P (1967) Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13(1): 21–27.

Craik FIM & Rose NS (2012) Memory encoding and aging: A neurocognitive perspective. Neuroscience and Biobehavioral Reviews 36(7): 1729–39.

Cullen JD, Athi N, Al-Jader M, Johnson P, Al-Shammaa AI, Shaw A & El-Rasheed AMA (2008) Multisensor fusion for on line monitoring of the quality of spot welding in automotive industry. Measurement 41(4): 412–423.

Devijver PA & Kittler J (1982) Pattern Recognition: A Statistical approach. Prentice Hall, Englewood Cliffs, NJ.

Dietterich TG (2002) Machine learning for sequential data: A review. Proc. Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, Springer, 15–30.

Duda RO, Hart PE & Stork DG (2001) Pattern Classification. John Wiley & Sons, New York, NY.

El Ouafi A, Belanger R & Guillot M (2012) Dynamic resistance based model for on-line resistance spot welding quality assessment. Materials Science Forum 706-709: 2925–2930.

Fiebrink R & Fujinaga I (2006) Feature selection pitfalls and music classification. Proc. 7th International Conference on Music Information Retrieval, 340–341.

Fitzmaurice GM, Laird NM & Ware JH (2011) Applied longitudinal analysis. John Wiley & Sons, Hoboken, NJ, 2nd edition.

Fredericks T, Choi S, Hart J, Butt S & Mital A (2005) An investigation of myocardial aerobic capacity as a measure of both physical and cognitive workloads. International Journal of Industrial Ergonomics 35(12): 1097–1107.

Freedson PS, Melanson E & Sirard J (1998) Calibration of the Computer Science and Applications, Inc. accelerometer. Medicine and Science in Sports and Exercise 30(5): 777–781.

French JW (1951) The Description of Aptitude and Achievement Tests in Terms of Rotated Factors. Psychometric Monographs. University of Chicago Press, Chicago, IL.

Friesen GM, Jannett TC, Jadallah MA, Yates SL, Quint SR & Nagle HT (1990) A comparison of the noise sensitivity of nine QRS detection algorithms. IEEE Transactions on Biomedical Engineering 37(1): 85–98.

Froberg K & Andersen LB (2010) The importance of physical activity for chidhood health. Proc. 5th International Congress of Youth Health, 41–46.

Gelman A & Pardoe I (2007) Average predictive comparisons for models with nonlinearity, interactions, and variance components. Sociological Methodology 37(1): 23–51.

Gevins AS & Rémond A (eds) (1987) Methods of Analysis of Brain Electrical and Magnetic Signals, volume I of *Handbook of Electroencephalography and Clinical Neurophysiology*. Elsevier, Amsterdam.

Grimes D, Tan DS, Hudson SE, Shenoy P & Rao RP (2008) Feasibility and pragmatics of classifying working memory load with an electroencephalograph. Proc. SIGCHI Conference on Human Factors in Computing Systems, 835–844.

Guyon I & Elisseeff A (2006) An introduction to feature extraction. In: Guyon I, Nikravesh M, Gunn S & Zadeh LA (eds) Feature Extraction: Foundations and Applications, volume 207 of *Studies in Fuzziness and Soft Computing*, 1–25. Springer Berlin Heidelberg.

Hand DJ, Mannila H & Smyth P (2001) Principles of data mining. MIT Press, Cambridge, MA.

Hart JD & Wehrly TE (1986) Kernel regression estimation using repeated measurements data. Journal of the American Statistical Association 81(396): 1080–1088.

Hart SG & Staveland LE (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research, 239–250. North Holland Press, Amsterdam.

Hastie T, Tibshirani R & Friedman J (2009) The elements of statistical learning : data mining, inference, and prediction. Springer Series in Statistics. Springer-Verlag, New York, NY, 2nd edition.

Hendelman D, Miller K, Baggett C, Debold E & Freedson P (2000) Validity of accelerometry for the assessment of moderate intensity physical activity in the field. Medicine and Science in Sports and Exercise 32(9): 442–449.

Holmström L, Koistinen P, Laaksonen J & Oja E (1997) Neural and statistical classifiers-taxonomy and two case studies. IEEE Transactions on Neural Networks 8(1): 5–17.

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24(6): 417.

Hyndman RJ & Athanasopoulos G (2014) Forecasting: principles and practice. OTexts.

Iqbal ST, Adamczyk PD, Zheng XS & Bailey BP (2005) Towards an index of opportunity: understanding changes in mental workload during task execution. Proc. SIGCHI Conference on Human Factors in Computing Systems, 311–320.

Jain AK, Duin RPW & Mao J (2000) Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(1): 4–37.

Kantardzic M (2011) Data mining: concepts, models, methods, and algorithms. Wiley-IEEE Press, New York, NY.

Knoll A, Wang Y, Chen F, Xu J, Ruiz N, Epps J & Zarjam P (2011) Measuring cognitive workload with low-cost electroencephalograph. Proc. 13th IFIP TC.13 International Conference on Human-Computer Interaction, 568–571.

Koenig A, Novak D, Omlin X, Pulfer M, Perreault E, Zimmerli L, Mihelj M & Riener R (2011) Real-time closed-loop control of cognitive load in neurological patients during robot-assisted gait training. IEEE Transactions on Neural Systems and Rehabilitation Engineering 19(4): 453–64.

Kohavi R & Sommerfield D (1995) Feature subset selection using the wrapper method: Overfitting and dynamic search space topology. Proc. 1st International Conference on Knowledge Discovery and Data Mining, 192–197.

Kohonen T, Hynninen J, Kangas J, Laaksonen J & Torkkola K (1996) LVQ PAK: The learning vector quantization program package. Technical report, Helsinki University of Technology.

Kudo M & Sklansky J (2000) Comparison of algorithms that select features for pattern classifiers. Pattern Recognition 33(1): 25–41.

Kugiumtzis D & Tsimpiris A (2010) Measures of analysis of time series (MATS): A MATLAB toolkit for computation of multiple measures on time series data bases. Journal of Statistical Software 33: 1–30.

LaPorte RE, Caspersen HJ & Montoye CJ (1985) Assessment of physical activity in epidemiologic research: Problems and prospects. Public Health Reports 100(2): 131–146.

Marsland S (2014) Machine learning: An algorithmic perspective. Chapman and Hall/CRC, Boca Raton, FL.

McGrew KS (2009) CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. Intelligence 37(1): 1–10.

Mehler B, Reimer B & Coughlin JF (2012) Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: An on-road study across three age groups. Human Factors 54(3): 396–412.

Mehler B, Reimer B, Coughlin JF & Dusek JA (2009) Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. Transportation Research Record: Journal of the Transportation Research Board 2138(1): 6–12.

Mitchell TM (1997) Machine Learning. McGraw-Hill, New York, NY.

Moore AW & Lee MS (1994) Efficient algorithms for minimizing cross validation error. Proc. 11th International Conference on Machine Learning, 190–198.

Mörchen F (2003) Time series feature extraction for data mining using DWT and DFT. Technical report no. 33. Philipps-University Marburg.

Narendra P & Fukunaga K (1977) A branch and bound algorithm for feature selection. IEEE Transactions on Computers C-26(9): 917–922.

NeuroSky (2009) ThinkGear API and Reference Manual.

Noel JB, Bauer KW & Lanning JW (2005) Improving pilot mental workload classification through feature exploitation and combination: A feasibility study. Computers & Operations Research 32(10): 2713–2730.

Oppenheim AV & Schafer RW (2010) Discrete-Time Signal Processing. Prentice-Hall, Inc., Upper Saddle River, NJ, 3rd edition.

Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philosophical Magazine 2(6): 559–572.

Peng Y, Wu Z & Jiang J (2010) A novel feature selection approach for biomedical data classification. Journal of Biomedical Informatics 43(1): 15–23.

Piquado T, Isaacowitz D & Wingfield A (2010) Pupillometry as a measure of cognitive effort in younger and older adults. Psychophysiology 47(3): 560–569.

Prabhu KMM (2013) Window functions and their applications in signal processing. CRC Press, Boca Raton, FL.

Pudil P, Novovicova J & Kittler J (1994) Floating search methods in feature selection. Pattern Recognition Letters 15(11): 1119–1125.

Pyle D (1999) Data preparation for data mining. Morgan Kaufmann Publishers, San Francisco, CA.

Raudys S (2006) Feature over-selection. In: Yeung DYY, Kwok JT, Fred A, Roli F & de Ridder D (eds) Structural, Syntactic, and Statistical Pattern Recognition, volume 4109 of *Lecture Notes in Computer Science*, 622–631. Springer, Heidelberg.

Reimer B, Mehler B, Coughlin JF, Godfrey KM & Tan C (2009) An on-road assessment of the impact of cognitive workload on physiological arousal in young adult drivers. Proc. 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 115–118.

Reinsch CH (1967) Smoothing by spline functions. Numerische Matematik 10: 177–183.

Reinsch CH (1971) Smoothing by spline functions, II. Numerische Matematik 16: 451–454.

Reunanen J (2003) Overfitting in making comparisons between variable selection methods. Journal of Machine Learning Research 3: 1371–1382.

Reunanen J (2004) A pitfall in determining the optimal feature subset size. Proc. 4th International Workshop on Pattern Recognition in Information Systems, 176–185.

Rothney MP, Neumann M, Béziat A, Chen KY & A (2007) An artificial neural network model of energy expenditure using non-integrated acceleration signals. Journal of Applied Physiology 103(4): 1419–1427.

Saari P, Eerola T & Lartillot O (2011) Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. IEEE Transactions on Audio, Speech, and Language Processing 19(6): 1802–1812.

Salthouse TA (2010) Major issues in cognitive aging. Oxford University Press, Oxford, NY.

Sato A & Yamada K (1996) Generalized learning vector quantization. Advances in neural information processing systems 423–429.

Schaffer C (1993) Selecting a classification method by cross-validation. Machine Learning 13(1): 135–143.

Seo S & Obermayer K (2003) Soft learning vector quantization. Neural computation 15(7): 1589–1604.

Smialowski P, Frishman D & Kramer S (2010) Pitfalls of supervised feature selection. Bioinformatics 26(3): 440–443.

Staudenmayer J, Pober D, Crouter S, Bassett D & Freedson P (2009) An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. Journal of Applied Physiology 107(4): 1300–1307.

Stearns SD (1976) On selecting features for pattern classifiers. Proc. 3rd International Conference on Pattern Recognition, 71–75.

Swartz AM & Strath SJ (2000) Estimation of energy expenditure using CSA accelerometers at hip and wrist sites. Medicine and Science in Sports and Exercise 32(9): 450–456.

Theodoridis S & Koutroumbas K (2008) Pattern Recognition. Academic Press, Boston, MA, 4th edition.

Troiano RP (2006) Translating accelerometer counts into energy expenditure: Advancing the quest. Journal of Applied Physiology 100(4): 1107–1108.

Verbeke G & Molenberghs G (2000) Linear Mixed Models for Longitudinal Data. Springer Series in Statistics. Springer, New York, NY.

Whitney AW (1971) A direct method of nonparametric measurement selection. IEEE Transactions on Computers 100(9): 1100–1103.

Wilson GF (2002) An analysis of mental workload in pilots during flight using multiple psychophysiological measures. International Journal of Aviation Psychology 12(1): 3–18.

Wolpert DH (1992) On the connection between in-sample testing and generalization error. Complex Systems 6(1): 47.

Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou ZH, Steinbach M, Hand DJ & Steinberg D (2008) Top 10 algorithms in data mining. Knowledge and Information Systems 14(1): 1–37.

Wylie N, Wylie SR, Cullen JD, Al-Jader M, Al-Shamma'a AI & Shaw A (2010) NDE system for the quality control of spot welding in the automotive industry. Proc. IEEE Sensors Applications Symposium, 73–78.

Xing Z, Pei J & Keogh E (2010) A brief survey on sequence classification. ACM SIGKDD Explorations Newsletter 12(1): 40–48.

Original publications

I Haapalainen E*, Laurinen P, Junno H, Tuovinen L & Röning J (2005) Methods for classifying spot welding processes: a comparative study of performance. Proc. 18th International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems, 412-421.

II Haapalainen E*, Laurinen P, Junno H, Tuovinen L & Röning J (2008) Feature selection for identification of spot welding processes. In: Andrade-Cetto J, Ferrier J-L, Pereira JD, Filipe J (eds) Informatics in Control Automation and Robotics, volume 15 of *Lecture Notes in Electrical Engineering*, 69-79. Springer, Heidelberg.

III Haapalainen E*, Laurinen P, Siirtola P, Röning J, Kinnunen H & Jurvelin H (2008) Exercise energy expenditure estimation based on acceleration data using the linear mixed model. Proc. IEEE International Conference on Information Reuse and Integration, 131-136.

IV Haapalainen E*, Kim S, Forlizzi J & Dey A (2010) Psycho-physiological measures for assessing cognitive load. Proc. 12th ACM International Conference on Ubiquitous Computing, 301-310.

V Ferreira E, Ferreira D, Kim S, Siirtola P, Röning J, Forlizzi J & Dey A (2014) Assessing real-time cognitive load based on psycho-physiological measures for younger and older adults. Proc. IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain, 39-48.

*Eija Ferreira (née Haapalainen)

Reprinted with kind permission from Springer Science+Business Media (Publications I and II), IEEE (Publications III and V), and Association for Computing Machinery, Inc. (Publication IV).

Original publications are not included in the electronic version of the dissertation.

525. Lu, Pen-Shun (2015) Decoding and lossy forwarding based multiple access relaying

526. Suopajärvi, Terhi (2015) Functionalized nanocelluloses in wastewater treatment applications

527. Pekuri, Aki (2015) The role of business models in construction business management

528. Mantere, Matti (2015) Network security monitoring and anomaly detection in industrial control system networks

529. Piri, Esa (2015) Improving heterogeneous wireless networking with cross-layer information services

530. Leppänen, Kimmo (2015) Sample preparation method and synchronized thermography to characterize uniformity of conductive thin films

531. Pouke, Matti (2015) Augmented virtuality : transforming real human activity into virtual environments

532. Leinonen, Mikko (2015) Finite element method and equivalent circuit based design of piezoelectric actuators and energy harvester dynamics

533. Leppäjärvi, Tiina (2015) Pervaporation of alcohol/water mixtures using ultra-thin zeolite membranes : membrane performance and modeling

534. Lin, Jhih-Fong (2015) Multi-dimensional carbonaceous composites for electrode applications

535. Goncalves, Jorge (2015) Situated crowdsourcing : feasibility, performance and behaviours

536. Herrera Castro, Daniel (2015) From images to point clouds : practical considerations for three-dimensional computer vision

537. Komulainen, Jukka (2015) Software-based countermeasures to 2D facial spoofing attacks

538. Pedone, Matteo (2015) Algebraic methods for constructing blur-invariant operators and their applications

539. Karhu, Mirjam (2015) Treatment and characterisation of oily wastewaters

540. Panula-Perälä, Johanna (2015) Development and application of enzymatic substrate feeding strategies for small-scale microbial cultivations : applied for *Escherichia coli*, *Pichia pastoris*, and *Lactobacillus salivarius* cultivations

Book orders:
Granum: Virtual book store
http://granum.uta.fi/granum/

UNIVERSITY OF OULU P.O. Box 8000 FI-90014 UNIVERSITY OF OULU FINLAND

ACTA UNIVERSITATIS OULUENSIS

UNIVERSITY of OULU
OULUN YLIOPISTO

ISBN 978-952-62-0900-5 (Paperback)
ISBN 978-952-62-0901-2 (PDF)
ISSN 0355-3213 (Print)
ISSN 1796-2226 (Online)