Utrecht University

Doctoral Thesis

Temporal Segmentation using Support Vector Machines in the context of Human Activity Recognition

Author:
R.Q. Vlasveld

Supervisor:
Dr. James Smith

A thesis submitted in fulfilment of the requirements for the degree of Master of Science

in the

Research Group Name
Department or School Name

October 2013

Declaration of Authorship

I, R.Q. Vlasveld, declare that this thesis titled, 'Temporal Segmentation using Support Vector Machines in the context of Human Activity Recognition' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"Thanks to my solid academic training, today I can write hundreds of words on virtually any topic without possessing a shred of information, which is how I got a good job in journalism."

Dave Barry

UNIVERSITY NAME (IN BLOCK CAPITALS)

Abstract

Faculty Name
Department or School Name

Master of Science

Temporal Segmentation using Support Vector Machines in the context of Human Activity Recognition

by R.Q. Vlasveld

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too...

Acknowledgements

The acknowledgements and the people to thank go here, don't forget to include your project advisor...

Contents

List of Figures

List of Tables

Acronyms

AIC Akaike Information Criterion. 9, 42

BIC Bayesian Information Criterion. 9, 42

CUSUM Cumulative Sum. 3, 4, 7–9, 37, 42

HMM Hidden Markov Model. 9, 50

ICSS Iterated Cumulative Sums of Squares. 4, 9, 36

i.d.d. independently and identically distributed. 52

KLIEP Kullback-Leibler Importance Estimation Procedure. 4, 5, 8, 37, 42

MLE Maximum Likelihood Estimates. 9

MOSUM Moving Sum. 3, 7

OCC One-Class Classifier. 42

OC-SVM One-Class Support Vector Machine. 18, 42, 44

PCA Principal Component Analysis. 7, 8

PDF Probability Density Function. 2, 3

QP Quadratic Programming. 16, 19

RBF Radial Base Function. 17, 20–22, 55

ROC Receiver-Operating Characteristic. 41

SV Support Vector. 20

SVC Support Vector Classifier. 51

SVDD Support Vector Data Description. vi, viii, 18, 20, 22, 41, 42, 51, 56

SVM Support Vector Machine. viii, 3, 5, 6, 14–21, 41, 44, 55, 56

ν-SVM ν-Support Vector Machine. vi, viii, 18–20

SWAB Sliding Window And Buttom-up. 6, 7

Symbols

a	distance	m
P	power	W (Js^{-1})
ω	angular frequency	$rads^{-1}$

For/Dedicated to/To my...

Chapter 1

Introduction

1.1 Outline

Not intented for the reader.

- Context of research (human activity recognition), real-world applications
- Current methods, wrapper vs. filter methods
- Problem statement with current filter methods (which follows from Chapter 3 which goes in-depth with methods).
- Purpose of this research. E.g. "Find a better algorithm for short-activity segmentation"
- Relate to real-world applications
- Outline for rest of the thesis

Chapter 2

Literature review

2.1 Outline

Not intented for the reader.

- Literature review about Temporal Segmentation (previous draft was more about classification)
- Consider methods for the context of filter-methods for classification
- Take a loot at 3-4 different kind of methods for change detection:
  - Introduction with a lot of techniques
  - Explain why look at a few
  - CUSUM - or other more traditional methods
  - Density-ratio estimation
  - Support Vector Machines (?) - if there are more sources about this
  - (Dimensionality reduction) –¿ probably not
  - *Not to much attention to all techniques, focus is on SVM*
- With each method, shortly look at characteristics, strengths and weaknesses and consider applicability to accelerometer sensor data

2.2 Statistical framework

Many applications require the detection of time points at which the underlying properties of a system change. This problem has received a lot of attention in the fields of data

mining, etc... *** list and refs ***. Often this problem is formulated in a statistical framework, by inspecting the underlying data generating Probability Density Function (PDF) of the time series data. A change point is then defined as a significant change in the properties of the PDF, such as the mean and variance.

The widely used Cumulative Sum (CUSUM) method by Basseville *et al.* [9] takes this approach. It originates from control methods for detection from bench marks. This method and some derivatiges are discussed and analyzed in section 2.3.

Many methods rely on pre-specified parametric model assumptions and considers the data be independent over time, which makes it less flexible to real-world applications. The methods proposed by Kawahara *et al.* [35] and Lui *et al.* [40] try to overcome these problems by estimating the *ratio* between the PDF, instead of estimating each PDF. This approach is discussed and analyzed in section 2.4.

The density-estimation methods rely on the log-likelihood ratio between PDFs. The method of Camci [13] takes an other approach within the statistical framework, by using a SVM. One problem it tries to overcome is the (claimed) weakness of many methods to detect a decrease in variance. The method represents the distribution over the data points as a hyper-sphere in a higher dimension using the kernel trick. A change in the PDF is represented by a change in the radius of this sphere. Section 2.6 discusses the SVM-method. *** Put emphasis that this is the source of inspiration for the chosen method? ***

*** Maybe add dimensionality reduction, but for now leave out ***

2.3 CUSUM

Notes: Non-Bayesian change detection algorithm (i.e. no prior distribution believe available for the change time). The CUSUM method is developed by Page [48] for the application of statistical quality control (it is also known as a control chart). Primary for detection of mean shift. The Moving Sum (MOSUM) of squares test for monitoring variance changes [31]. Use of Cumulative Sums of Squares for Retrospective Detection of Changes of Variance [33]

An often used approach in the statistical framework of change detection is the CUSUM as introduced by Page [48]. Originally used for quality control in production environments, its main function is to detect change in the mean of measurements and has been applied to this problem [9]. It is an non-Bayesian method and thus makes no assumptions (or: prior belief distributions) for the change points. Many extensions to this method have

been proposed. Some focus on the change in mean, such the method of Alippi and Roveri [4]. Others apply the method the problems in which the change of variance is under consideration. Among others are there the centered version of the cumulative sums, introduced by Brown, Durbin and Evans [12] and the MOSUM of squares by [31].

The method of Inclán and Tiao [33] builds on the centered version of CUSUM [12] to detect changes in variance. Using the Iterated Cumulative Sums of Squares (ICSS) algorithm they are able to find multiple change points in a reflective manner. Let $C_k = \sum_{i=1}^{k} \alpha_t^2$ be the cumulative sum of squares for a series of uncorrelated random variables $\{\alpha_t\}$ of length T. The centered (and normalized) sum is squares is defined as

$$D_k = \frac{C_k}{C_T} - \frac{k}{T}, \quad k = 1, \ldots, T, \quad \text{with } D_0 = D_T = 0. \tag{2.1}$$

For a series with homogeneous variance, the value of D_k will oscillate around 0. In case of a sudden change, the value will increase and exceed some predefined boundary with high probability. The behavior of D_k is related a Brownian bridge. By using an iterative algorithm, the method is able to minimize the masking effect of successive change points.

One of the motivations for the ICSS algorithm was the heavy computational burden involved with Bayesian methods, which need to calculate the posterior odds for the log-likelihood ratio testing. The ICSS algorithm avoids applying a function at all possible locations, due to the iterative search. The authors recommend the algorithm for analysis of long sequences.

2.4 Change-detection by Density-Ratio Estimation

Many approaches to detect change points, regarded as a change in the underlying probabilistic generation, monitor the logarithm of the likelihood ratio between two consecutive intervals. Some methods which rely on this are novelty detection, maximum-likelihood estimation and online learning of autoregressive models [35]. A limitation of these methods is that they rely on pre-specified parametric models. Non-parametric models for density estimation have been proposed, but it is said to be a hard problem [27, 53]. A solutions to this is to estimate the *ratio* of probabilities instead of the probabilities themselves. One of the recent methods to achieve this is the Kullback-Leibler Importance Estimation Procedure (KLIEP) by Sugiyama *et al.* [54].

The method proposed by Kawahara and Sugiyama [35] is composed of an online version of the KLIEP algorithm. The method also considers *sequences* of samples (rather than samples directly) because the time series samples are generally not independent over time. An advantage over other non-parametric approaches, such as sequential one-class

support vector machines, is that the model has an natural cross-validation procedure. This makes that the value of tuning parameters, such as the kernel bandwidth, can be objectively obtained.

In their formulation change is detected by monitoring the logarithm of the likelihood ratio between the reference (past) and test (current) time intervals

$$S = \sum_{i=1}^{n_{\text{te}}} \ln \frac{p_{\text{te}}(\boldsymbol{Y}_{\text{te}}(i))}{p_{\text{rf}}(\boldsymbol{Y}_{\text{te}}(i))} \tag{2.2}$$

Where $\boldsymbol{Y}_{\text{te}}(i)$ is a sequence of samples from the test interval. A change is detected when $S > \mu$, for some predetermined threshold μ. The question is then how to calculate the density ratio

$$w(\boldsymbol{Y}) := \frac{p_{\text{te}}(\boldsymbol{Y}))}{p_{\text{rf}}(\boldsymbol{Y})} \tag{2.3}$$

because this ratio is unknown and should be estimated. The naive approach is to estimate the ratio by taking the ratio of the estimated densities. Since this is known to be a hard problem and sensitive for errors, the solution would be to estimate the ratio directly.

The procedure of the method proposed by Kawahara and Sugiyama [35] is to first apply the batch KLIEP algorithm with model selection for initial parameter $\boldsymbol{\alpha}$ and kernel width calculation. Then for every new sample the reference and test intervals are shifted and the calculated parameters $\boldsymbol{\alpha}$ are updated. Finally the logarithm of the likelihood ratio is evaluated. If it is above the threshold μ the current time is reported as a change point.

Improvements in this line of research, by Liu *et al.* [40] has led the application of improved density-ratio estimation methods to the problem of change detection. Such an improvement is the Unconstrained Least-Squares Importance Fitting (uLSIF) method [34] and an extension which possesses a superior non-parametric convergence property: Relative uLSIF (RuLSIF) [64].

2.5 Temporal Segmentation

- Given overview of segmentation techniques, for times series data
- Use different "point-of-views", or terms
- "Segmentation"
- "Change detection"
- "Novelty detection"

- Specific view on SVMs

This section gives an overview of the literature on temporal segmentation in the context of human activity recognition. It takes a look on different implementations and methodologies. A wide range of terms and subtle differences are used in the field, such as 'segmentation', 'change detection', 'novelty detection' and 'outlier detection'. These will be the categorical terms for which we discuss the literature. Finally we will discuss other applications of SVMs in the context of these terms.

2.5.1 Segmentation

*** This subsection is mainly from previous draft version ***

*** TODO: Create compact notation, one sentence per paper max ***

Segmentation methods can roughly be categorized in three methods in the way the data is processed, as discussed by Avci *et al.* [6]:

- **Top-down** methods iteratively divide the signal in segments by splitting at the best location. The algorithm starts with two segments and completes when a certain condition is met, such as when an error value or number of segments k is reached. These methods process the data points recursively, which results in a complexity of $O(n^2k)$.
- **Bottom-up** methods are the natural complement to top-down methods. They start with creating $n/2$ segments and iteratively join adjacent segments while the value of a cost function for that operation is below a certain value. Given the average segment length L, the complexity of this method is $O(nL)$.
- **Sliding-window** methods are simple and intuitive for online segmenting purposes. It starts with a small initial subsequence of the time series. New data points are joined in the segment until the fit-error is above a threshold. Since the data is only processed very locally, these methods can yield in poor results [37]. The complexity is equal to the bottom-up approach, $O(nL)$, where L is the average segment length.
- **Sliding Window And Buttom-up**, as introduced by Keogh *et al.* [37], joins the ability of the sliding window mechanism to process time series online and the bottom-up approach the create superior segments in terms of fit-error. The algorithm processes the data in two stages. The first stage is to join new data points in the current segment created by a sliding window, and pass this to a

buffer with space for a few segments. The buffer then processes the data Bottom-up and returns the first (left-most) segment as final segment. Because this second stage retains some (semi-)global view of the data, the results are comparative with normal Bottom-up. It is stated by Keogh *et al.* that the complexity of Sliding Window And Buttom-up (SWAB) is a small constant factor worse than that of regular Bottom-up.

It is clear that for the application of this research sliding-window and preferably SWAB-based algorithms should be considered.

The SWAB method proposed by Keogh *et al.* [37] is dependent on an user setting, providing the maximum error when performing both stages. Each segment is approximated by using piecewise linear representation (PLR), an often used method. The user provided error threshold controls the granularity and number of segments. Other methods have been proposed, such as an adaptive threshold based on the signal energy by Guenterberg *et al.* [23], the adaptive CUSUM-based test by Alippi *et al.* [4] and the MOSUM by Hsu [31] in order to eliminate this user-dependency. The latter of these methods is able to process the accelerometer values directly, although better results are obtained when features of the signal are processed, as done in the former method. Here the signal energy, mean and standard deviation are used to segment activities and by adding all the axial time series together, the Signal-To-Noise ration is increased, resulting in a robuster method.

The method of Guenterberg *et al.* extracts features from the raw sensor signal to base the segmentation on other properties than the pure values. The method of Bernecker *et al.* [11] uses other statistical properties, namely autocorrelation, to distinguish periodic from non-periodic segments. Using the SWAB method the self-similarity of a one-dimensional signal is obtained. The authors claim that only a slight modification is needed to perform the method on multi-dimensional data. After the segmentation phase, the method of Bernecker *et al.* extracts other statistical features which are used in the classification phase.

The proposal of Guo *et al.* [24] dynamically determines which features should be used for the segmentation and simultaneously determines the best model to fit the segment. For each of the three dimensions features such as the mean, variance, covariance, correlation, energy and entropy are calculated. By extending the SWAB method, for every frame a feature set is selected, using an enhanced version of Principal Component Analysis (PCA). The research also considered the (Stepwise) Feasable Space Window as introduced by [41], but since it results in a higher error rate than SWAB the latter was chosen to extend. Whereas the before mentioned algorithms use a linear representation, this

methods considers linear, quadratic and cubical representations for each segment. This differs from other methods where the model is fixed for the whole time series, such as [21], which is stated to perform inferior on non-stationary time series such as daily life.

The time series data from a sensor can be considered as being drawn from a certain stochastic process. Probabilistic models can be constructed on that signal, yielding in probabilistic and Bayesian based segmentation methods. The CUSUM-methods takes a statistical approach and relies on the log-likelihood ratio [26] to measure the difference between two distributions. To calculate the ratio, the probability density functions need to be calculated. The method of Kawahara *et al.* [35] proposes to estimate the ratio of probability densities (known as the *importance*), based on the log likelihood of test samples, directly, without explicit estimation of the densities. The method by Liu *et al.* [40] uses a comparable dissimilarity measure using the KLIEP algorithm. They claim this results in a robuster approach for real-world scenarios. Although this is a model-based method, no strong assumptions (parameter settings) are made on the models.

The method of Adams and MacKay [1] builds a probabilistic model on the segment run length, given the observed data so far. Instead of modeling the values of the data points as a probabilistic distribution, the length of segments as a function if time is modeled by calculating the posterior probability. It uses a prior estimate for the run length and a predictive distribution for newly-observed data, given the data since the last change point. This method contrasts with the approach of Guralnik and Srivastava [25] in which change points are detected by a change in the (parameters of an) underlying, observed, model. For each new data point, the likelihoods of being a change point and part of the current segment are calculated, without a prior model (and thus is a non-Bayesian approach). It is observed that when no change point is detected for a long period of time, the computational complexity increases significantly.

Another application of PCA is to characterize the data by determining the dimensionality of a sequence of data points. The proposed method of Berbič *et al.* [7] determines the number of dimensions (features) needed to approximate a sequence within a specified error. With the observation that more dimensions are needed to keep the error below the threshold when transitions between actions occur, cut-points can be located and segments will be created. The superior extension of their approach uses a Probabilistic PCA algorithm to incorporate the dimensions outside the selected set as noise.

In the method by [29] a cost function is defined over segments of data which is to be minimized. The cost functions thereby searches for internally homogeneous segments of data, reflecting states in which the devices and the user are. The cost function can be any arbitrary function and in the implementation the sum of variances over the segments is used. Both in a local and global iterative replacement procedure (as an alternative

for the computationally hard dynamic programming algorithm) the best breakpoint locations c_i for a pre-defined number of segments $1 \leq i \leq k$ are optimized.

Many methods obtain an implicit segmentation as a result of classification over a sliding window *** add refs ***. The method of [65] explicitly performs segmentation and classification simultaneously. It argues that the classification of a pre-segmented test-sequences becomes straightforward with many classical algorithms to choose from. The algorithm matches test examples with the *sparsest* linear representation of mixture subspace models of training examples, searching over different temporal resolutions.

The method of Chamroukhi *et al.* [15] is based on a Hidden Markov Model (HMM) and logistic regression. It assumes a K-state hidden process with a (hidden) state sequence, each state providing the parameters (amongst which the order) for a polynomial. The order of the model segment is determined by model selecting, often using the Bayesian Information Criterion (BIC) or the similar Akaike Information Criterion (AIC) [3], as in [28].

Field of computer vision: [68], [39].

2.5.2 Change detection

*** change order of this and temporal segmentation sections? So first change, then segmentation? ***
Whereas the above mentioned researches focus on *segmentation*, many have focused on *change detection*. Although these techniques are closely related, there is a subtle difference. In the case of *change detection* to goal is to find, possibly unrelated, sudden change points in a signal [56]. In contrast, the goal of *temporal segmentation* is to find homogeneous segments of data, which can be the result of multiple detected changes.

The ICSS by Inclán and Tiao [33] is a statistical method which obtains comparable results (when applied to stock data) compared to Maximum Likelihood Estimates (MLE) and Bayesian. Whereas CUSUM can be applied to search for a change in mean, the ICSS is adapted to find changes in variance. It obtains a *likelihood ratio* for testing the hypothesis of one change against no change in the variance. Using an iterative approach, all possible change points are considered. The proposal of [17] extends on the CUSUM-based methods to find change points in mean and variance, by creating a more efficient and accurate algorithm.

*** TODO: move CUSUM based techniques to this subsection ***

— Segmentation —

"Segmentation and Recognition of Motion Streams by Similarity Search" [39]. 29, 2007

"Aligned Cluster Analysis for Temporal Segmentation of Human Motion" [68]. 63, 2008

— Change detection —

"A unifying framework for detecting outliers and change points from time series" [56]. 87, 2006

"Sequential change-point detection based on direct density-ratio estimation" [36]. 22, 2012

"Change point detection in time series data using support vectors" [13]. 3, 2010

— Novelty detection —

"Online novelty detection on temporal sequences" [42]. 146, 2003
"Time-series novelty detection using one-class support vector machines" [43]. 78, 2003

"Novelty detection: a reviewpart 1: statistical approaches" [44]. 697, 2003

"Support Vector Method for Novelty Detection" [50]. 337, 1999

— Outlier detection — "A unifying framework for detecting outliers and change points from time series" [56]. 87, 2006

"Outliers in statistical data" [8] (book). 3745, 1994

"A survey of outlier detection methodologies" [30]. 791, 2004

"Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection" [51]. 2, 2012

— SVMs —

"Change point detection in time series data using support vectors" [13]. 3, 2010

"Time-series novelty detection using one-class support vector machines" [43]. 78, 2003

"Support Vector Method for Novelty Detection" [50]. 337, 1999

"Support vector domain description" [59]. 907, 1999 "Support vector data description applied to machine vibration analysis" [57]. 76, 1999

"Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support Vector Machine" [5]. 13, 2012

"A geometric approach to support vector machine (SVM) classification" [45]. 136, 2006

"Support vector machines in remote sensing: A review" [46]. 150, 2011

"Predicting time series with support vector machines" [47]. 693, 1997

2.6 Change-detection by Support Vector Machines

Introduced by Cortes and Vapnik [61, 62], Support Vector Machines offer a way to segment, and classify, linear separable data. This technique can also be applied to estimate density functions of given time series [63]. When combined with a mapping function, which maps the data from the input space I to a higher dimension feature space F, the input data can be non-linear separable. A separating linear hyperplane that segments the data in the feature space F, yields a non-linear segmentation in the lower-dimensional input space I. Instead of explicitly mapping the input data to the higher dimensional space, a kernel function can be used. This kernel function can calculate values of the feature space directly, without the need to first map the input values to this space. This process is referred to as the kernel trick.

The proposed method of Camci [13] uses a one-class support vector machine to segment time series data. One-class SVMs are used to describe the current data under consideration, by assuming all data points are from the same class [58]. To cope with possible

errors or outliers a soft-margin is applied [19]. The class is described by a spherical boundary around the data with center c and radius r, such that the volume is minimized. New data points are consecutively mapped from the input space to the feature space. The retrieved (high dimension) data point can be in- or outside the earlier constructed hyper-sphere, thereby giving information about a possible change point. *** NOTE: using the in/out position of a new datapoint is different from the approach in this thesis, by only using the model properties. *** Following the definition of Camci [13], the class description is obtained by minimizing r^2:

$$\text{Min } r^2 \tag{2.4}$$

$$\text{Subject to : } \|\mathbf{x}_i - \mathbf{c}\|^2 \leq r^2 \ \forall i, \ \mathbf{x}_i : i\text{th data point} \tag{2.5}$$

To be able to handle outliers in the input data, a penalty cost function ε_i for each outlier can be added.

Using this one-class SVM formulation, differences between two (consecutive) windows of data points with size w can be obtained. The first window is used as the input set, h_1 and the second as the test set h_t. For the first window a one-class SVM is constructed, yielding in a representation by c_1 and r_1. When the data points of the second window belong to the same class, the representation of that one-class SVM would equal the first:

$$c_1 = c_2, \quad r_1 = r_2 \tag{2.6}$$

In case the second window of data points does not belong to the same class, i.e. the probability density function that describes the data differs from the first, the describing values of the second window will differ from the first. A difference in the SVM center c or radius r represent a change in the mean and variance, respectively. The amount of difference can be expressed by a dissimilarity measure over the representations. When the dissimilarity between the two windows exceeds some predefined threshold th, there exists a change point between the windows.

*** Give visual explanation with circles and in- and out-side new data points ***

*** Make clear what approach we use (model properties) and create link/bridge to next Chapter (3), which discusses SVM in detail. ***

Chapter 3

Change detection by Support Vector Machines

3.1 Outline

Not intented for the reader.

- Content follows blogpost, but more formal
- Explain the connection between change detection and {outlier detection, density estimation, etc. }
- Explain change detection in relation with "M-Estimators".
- Explain why and how SVM can be used for outlier detection, and thus for change detection in time series
- Explain options when using SVM, such as RBF kernel
- Relate characteristics of (accelerometer) data to the used (RBF?) kernel
- Define some quality metrics –¿ this should be in a different chapter (About experiments)
- Note: try to avoid the perspective of *density estimation*. Instead: *data description* or *boundary description*.

*** TODO: *** Follow methodology of "A unifying framework for detecting outliers and change points from time series" [56]. It creates a two-stage process of first searching for outliers, and then using the "outlier-score" to find (sudden) change points, by a weighted

average of a moving window. That eventual score can be thresholded (as in the paper) or processed with something like CUSUM (proposal). Looks like my proposed methods, in that is combines outliers and gives a score to change points.

3.2 Outlier and change detection

3.2.1 General framework for outlier detection

In [51] a general framework for outlier detection is proposed. The authors compare existing methods and identified the common building block of the algorithms. The focus is on unsupervised methods, using information from a local selection of data objects for the detection of outliers. The following common algorithmic building blocks are identified:

1. Context: a "local" context of an object o for model building
2. Model: the method used for building the model
3. Reference: a "reference" context of object o for model comparison
4. Comparison: the method used for model comparison
5. Normalization: a (global) normalization procedure

Here the *context* and *reference* are sets of objects used for model building and model comparison, respectively.

3.2.2 Regression, Classification, etc

General framework of outlier detection, change detection in context of (simple?) regression, classification.

3.2.3 M-estimators

*** Move this section to somewhere in the end; is specification/extension, not base material *** To make a method less sensitive to outliers (in the training data) techniques from robust statistics are applied. The term *robustness* has many interpretations, one of them that it "signifies insensitivity to small deviations from the [prior] assumptions

[about the underlying situation]", according to Huber [32]. Methods have been proposed to apply robust statistics in the form of M-estimators to SVMs, such as [16, 18, 55].

*** Question: do slack-variables (allowance of outliers) make SVM robust? ***

3.3 One-Class Support Vector Machine

3.3.1 Support Vector Machine

We will first discuss the traditional two-class SVM before we consider the one-class variant, as introduced by Cortes and Vapnik in [19]. Consider a data set $\Omega = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$; points $x_i \in \mathbb{R}^d$ in a (for instance two-dimensional) space where x_i is the i-th input data point and $y_i \in \{-1, 1\}$ is the i-th output pattern, indicating the class membership.

A SVM can create a boundary between linear-separable data points, making it a non-probabilistic binary linear classifier. More flexible non-linear boundaries can be obtained by the use of a non-linear function $\phi(x)$, as illustrated in Figure 3.2. This function maps the input data from space $\mathcal{I}$ to a higher dimensional space $\mathcal{F}$. The SVM can create a linear separating hyperplane in the space $\mathcal{F}$ that separates the data points from the two classes. When the hyperplane is projected to the (lower) original input space $\mathcal{I}$ it creates a non-linear separating curve. The mapping and projection of data points can be efficient (and implicit) performed by using the kernel trick, which is discussed in section 3.3.2.

The separating hyperplane is represented by

$$w^T x + b = 0, \tag{3.1}$$

with $w \in F$ and $b \in R$. The hyperplane that is created determines the *margin* between the classes; the minimal distance from one of the data points to the hyperplane. In geometric sense, w is the normal vector to the hyperplane and $\frac{b}{\|w\|}$ determines the offset of the hyperplane to the origin. Since the distance between the two margins is equal to $\frac{2}{\|w\|}$, the maximum-margin hyperplane is found by minimizing $\|w\|$. The data points which lie on the margin are the *support vectors*. This geometrical interpretation is illustrated in Figure 3.1. All data points for which $y_i = -1$ are on one side of the hyperplane and all other data points (for which $y_i = 1$) are on the other side. The minimal distance from a data point to the hyperplane is for both classes equal. This results in a *maximal margin* between the two classes. Thus, the SVM searches for a maximal separating hyperplane.

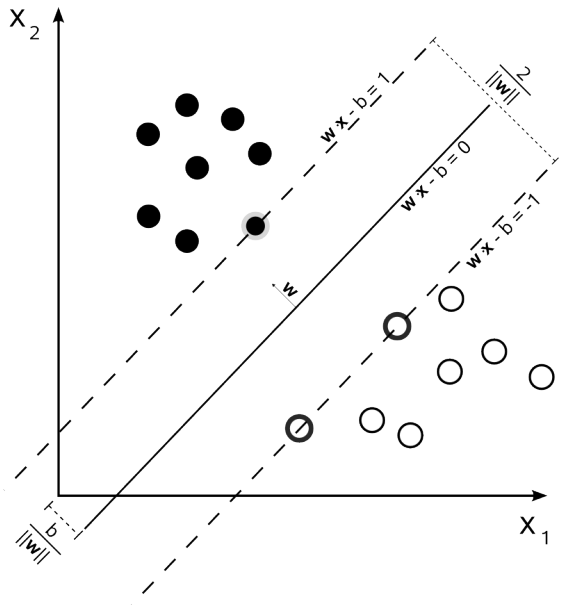

FIGURE 3.1: Illustration of the separating hyperplane of a SVM. Here w is the normal vector for the separating hyperplane and the distance between the two margins is $\frac{2}{\|w\|}$. Image from Wikipedia.org[1]

With every classification method there is a risk of overfitting. In that case the random error or noise of the data set is described instead of the underlying data. The SVM classifier can use a *soft margin* by allowing some data points to lie within the margin, instead of on the margin or farther away from the hyperplane. For this it introduces *slack variables* ξ_i for each data point and the constant $C > 0$ the determines the trade-off between maximizing the margin and the number of data points within that margin (and thus the training errors). The objective function for a SVM is the following minimization function:

$$\min_{w,\ b,\ \xi_i} \frac{\|w\|^2}{2} + C \sum_{i=1}^{n} \xi_i \tag{3.2}$$

subject to:

$$\begin{aligned} y_i(w^T \phi(x_i) + b) &\geq 1 - \xi_i \quad \text{for all } i = 1, \ldots, n \\ \xi_i &\geq 0 \quad \text{for all } i = 1, \ldots, n \end{aligned} \tag{3.3}$$

*** TODO: better format of above formula ***

This minimization problem can be solved (using Quadratic Programming (QP)) and transformed to its Lagrange dual formulation. To do so, the Lagrange multipliers $a_i >=$

[1] http://en.wikipedia.org/wiki/File:Svm_max_sep_hyperplane_with_margin.png

0 are introduced and the decision function becomes:

$$f(x) = \text{sgn}(\sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b), \tag{3.4}$$

where $K(x, x_i) = \phi(x)^T \phi(x_i)$ (which is further discussed in section 3.3.2). Here every data point $\mathcal{I}$ for which $a_i > 0$ is weighted in the decision function and thus "supports" the classification machine: hence the name "Support Vector Machine". Since it is shown that under certain circumstances SVMs show an equality to sparse representations [22, 52], there will often be relatively few Lagrange multipliers with a non-zero value.

3.3.2 Kernel trick

In the previous section, 3.3.1, the mapping function $\phi(x)$ and the kernel function K were briefly mentioned. The decision function in equation 3.4 only relies on the dot-products of mapped data points in the feature space $\mathcal{F}$ (i.e. all pairwise distances between the data points in that space). It shows [20] that as long as any function has the same result, without an explicit mapping to the higher dimension $\mathcal{F}$, the dot-products can be substituted by the kernel function K. This is known as the *kernel trick* and gives the SVM the ability to create non-linear decision function without high computational complexity. This mapping is illustrated in Figure 3.2. Here the non-linear separating boundary in the input space $\mathcal{I}$ is mapped, via ϕ, to a linear boundary in the feature space $\mathcal{F}$.

The kernel function K can have different forms, such as linear, polynomial and sigmoidal but the most used (and flexible) form is the Gaussian Radial Base Function (RBF). As Smola *et al.* [52] state, this Gaussian kernel yields good performance, especially when no assumptions can be made about the data. The kernel maps input space $\mathcal{I}$ to the feature space $\mathcal{F}$ which is a Hilbert Space of infinite dimensions *** ref needed ***:

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right), \tag{3.5}$$

where $\sigma \in R$ is a kernel parameter and $\|x - x'\|$ is the dissimilarity measure expressed in Euclidean distance.

3.3.3 Reproducing kernel Hilbert space

Write something on this topic? Or leave it implicit?

[2] http://en.wikipedia.org/wiki/File:Kernel_Machine.png

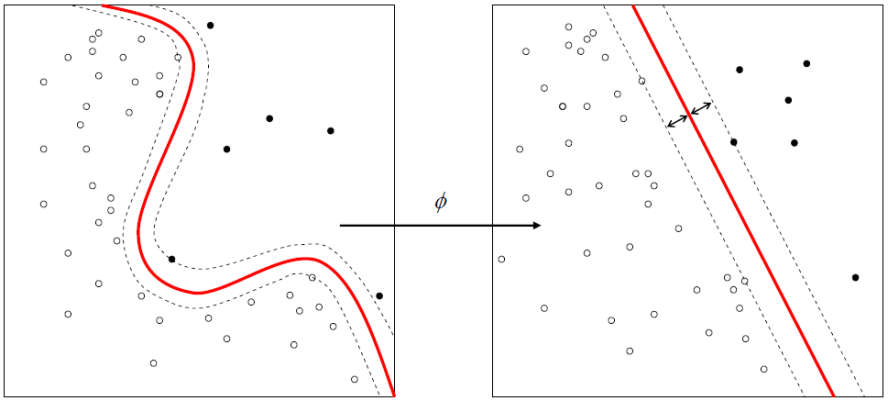

FIGURE 3.2: The non-linear boundary in the input space $\mathcal{I}$ (left) is transformed to a linear boundary in the feature space $\mathcal{F}$ (right) by mapping the data points with the function ϕ. The kernel trick uses a function K which performs an implicit mapping. Image from Wikipedia.org[2]

3.3.4 One-Class Support Vector Machines

The classical implementation of a SVM is to classify a dataset in two distinct classes. This is a common usecase, although sometimes there is no training data for both classes available. Still, one would like to classify new data points as regular, or in-class, or out-of-class, e.g. in the case of a novelty detection. With that problem only data from one class is available and the objective is to recognize new data points that are not part of that class. This problem is closely related to density estimation. In that context, the problem can be the following. Given an underlying probabilty distribution P the goal is to find a subset S for which the probability that a data point from P lies outside S equals some predetermined value ν between 0 and 1 [50].

We will discuss two implementations of One-Class Support Vector Machine (OC-SVM)s. The first is the ν-SVM by Schölkopf *et al.* [50] and closely follows the above problem statement regarding density estimation. The second, on which we will base our change detection method, is the SVDD by Tax and Duin [60].

3.3.5 ν-Support Vector Machines

The first of the OC-SVM methods we will discuss is often referred to as ν-SVM and introduced by Schölkopf *et al.* [50]. Instead of estimating the density function of an distribution P, it focuses on an easier problem: the algorithm find regions in the input where the "probability density lives". This results in a function such that most of the data is in the region where the function is nonzero.

The constructed decision function $\mathcal{F}$ resembles the function discussed in Section 3.3.1. It returns the value +1 in a (possibly small) region capturing most of the data points, and

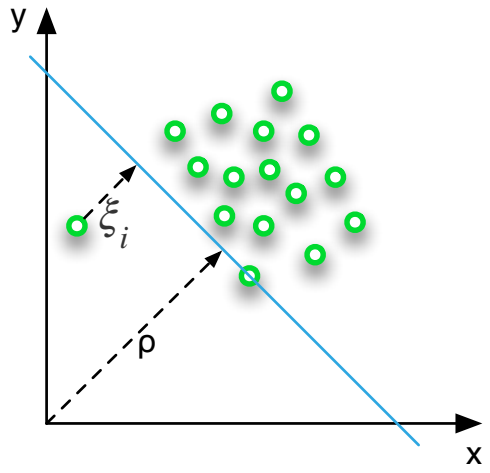

FIGURE 3.3: Graphical representation of ν-SVM. The separating hyperplane $w \cdot \phi(x_i) + \rho = 0$ create a maximal margin, in the feature space, between the data points and the origin. Slack variables ξ_i are used to create a soft margin.

-1 elsewhere. The method maps the data points from input space $\mathcal{I}$ to a feature space $\mathcal{F}$ (following classical SVMs). In that space $\mathcal{F}$ it separates the data points with maximal margin from the origin, with a hyperplane. For a new data points x, the function value $f(x)$ determines wheter the data point is part of the distribution (i.e. the value is $+1$) or a novelty (i.e. the value is -1). The hyperplane is represented by $g(x) = w \cdot \phi(x) + \rho = 0$ and the decision function is $f(x) = \text{sgn}(g(x))$. This hyperplane and the separation from the origin is illustrated in Figure 3.3.

The objective function to find the separating hyperplane is the following minimization function, which can be solved using QP:

$$\min_{w,\ \xi_i,\ \rho} \frac{\|w\|^2}{2} + \frac{1}{\nu n} \sum_{i=1}^{n} \xi_i - \rho \tag{3.6}$$

subject to:

$$\begin{aligned} (w \cdot \phi(x_i)) \geq & \rho - \xi_i \quad \text{for all } i = 1, \ldots, n \\ & \xi_i \geq 0 \quad \text{for all } i = 1, \ldots, n \end{aligned} \tag{3.7}$$

The decision function in the dual formulation with Lagrange multipliers is denoted as:

$$f(x) = \text{sgn}((w \cdot \phi(x_i)) - \rho) = \text{sgn}(\sum_{i=1}^{n} \alpha_i K(x, x_i) - \rho) \tag{3.8}$$

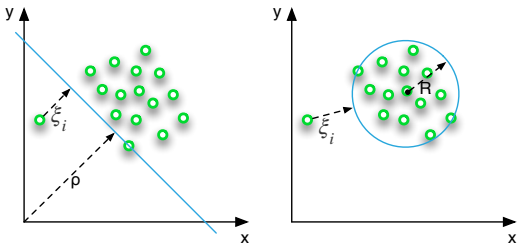

FIGURE 3.4: Graphical representation of the difference between ν-SVM (left) and SVDD (right). Note that for the sake of simplicity the kernel functions are not applied.

In the classical SVM objective function, as denoted in Equation 3.2, the parameter C decided the smoothness of the boundary, with respect to the slack variables ξ_i. In the formulation of ν-SVM the equivalent parameter is $\nu \in (0, 1)$ (hence the name). It characterizes the solution in two ways:

1. ν is an upper bound on the fraction of outliers, i.e. training examples regarded as out-of-class.
2. ν is a lower bound on the fraction of Support Vector (SV)s, i.e. training examples with a nonzero Lagrange multiplier α_i.

When ν approaches 0, the penalty factor for nonzero Lagrange multipliers ($\frac{1}{\nu n}$) becomes infinite, and thus the solution resembles a *hard margin* boundary.

This method creates a *hyperplane*, characterized by w and ρ, that separates the data with maximal margin from the origin in the feature space $\mathcal{F}$. In the following section we will discuss an alternative method, which uses an circumscribing *hypersphere* to characterize the data (and the region of distribution P where the density (or: support) lives).

3.3.6 Support Vector Data Description

The method introduced by Tax and Duin [60], known as Support Vector Data Description, takes as spherical instead of planar approach. The boundary, created in feature space $\mathcal{F}$, forms a hypersphere around the (high density region of the) data. The volume of this hypersphere is minimized to get the smallest enclosing boundary. By allowing outliers using slacks variables, in the same manner as classical SVM and ν-SVM, a soft margin is constructed.

The constructed hypersphere is characterized by a center $\mathbf{a}$ and a radius $R > 0$ as distance from the center to (any data point that is a SV on) the boundary, for which the volume, and thus the radius R, will be minimized. The center $\mathbf{a}$ is a linear combination

of of the support vectors. Like the classical SVM and SVDD it can be required that all the distances from the data points x_i to the center $\mathbf{a}$ are strict less then R (or equivalent measure, to create a hard margin) or soft margins are allowed by using slack variables ξ_i. In the case of a soft margin, the penalty is determined by C and the minimization is expressed as Equation 3.9. This principle is illustrated in the right image or Figure 3.4. Instead of a separating hyperplane, constructed by ν-SVM and illustrated on the left of the Figure, the SVDD creates a hypersphere (in the illustration a cricle) around the data points. By using kernel functions (e.g. the RBF) the hyperspheres in the high dimensional feature space $\mathcal{F}$ corresponds to a flexible and tight enclosing boundary in input space $\mathcal{I}$, illustrated in Figure 3.5.

$$\min_{R,\ \mathbf{a}} R^2 + C \sum_{i=1}^{n} \xi_i \tag{3.9}$$

subject to:

$$\begin{aligned} \|x_i - \mathbf{a}\|^2 \leq & R^2 + \xi_i \quad \text{for all } i = 1, \ldots, n \\ & \xi_i \geq 0 \quad \text{for all } i = 1, \ldots, n \end{aligned} \tag{3.10}$$

The dual formulation, by using Lagrange multipliers α_i, can be used to test a new data point z and is the following:

$$\|z - \mathbf{x}\|^2 = \sum_{i=1}^{n} \alpha_i \exp\left(\frac{-\|z - x_i\|^2}{\sigma^2}\right) \geq -R^2/2 + C_R \tag{3.11}$$

3.3.7 Kernel Function

- Explain RBF kernel
- Explain RKHS? (Hilbert Space)

$k(\mathbf{x}_i, \mathbf{x})$ is a kernel function computing a dot product in feature space, introduced by Aizerman *et al.* [2]. "Training a SVM with Gaussian RBF kernels corresponds to minimizing the specific cost function with a regularization operator" [52].
"Gaussian kernels tend to yield good performance under general smoothness assumptions and should be considered especially of no additional knowledge of the data is available." [52].
"Choosing a small width of the kernels leads to high generalization error as it effectively decouples the separate basis functions of the kernel expansion into very localized

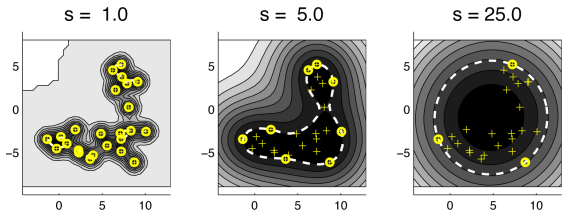

FIGURE 3.5: The SVDD method trained on a banana-shaped data set with different sigma-values for the RBF kernel. Solid circles are support vectors, the dashed line is the boundary. Image by Tax [58].

functions which is equivalent to memorizing the data, whereas a wide kernel tends to oversmooth." [52].

sparsity, RKHS: [22]

*** Maybe add section about "Change indication" rather than "detection"? ***

3.4 SVM parameters

3.5 Accelerometer-data characteristics

3.6 Quality metrics

*** *This section should be in the chapter about experiments* ***

Chapter 4

Proposed method

4.1 Outline

Not intented for the reader.

- Based on the problem statement with current research as stated in Chapter 3
- Adjust method to needs
- Explain using graphs, pseudo-algorithms. Make clear distinction in origin of ideas and why to apply

Chapter 5

Result

5.1 Outline

Not intented for the reader.

- Compare proposed method with methods of Chapter 2
- Provide plots, tables, graphs, error rates, precision, etc.
- Apply to a multiple of data, to compare to previous research - use that data
- Give theoretical analysis about performance. Big-O, memory, run-time, precision.
- This sections needs programmed implementations of own method and the ones compared

Chapter 6

Real-world applications

6.1 Outline

Not intented for the reader.

- Apply proposed method to real-world applications, such as
  - Daily life activity recognition (as the original context of this thesis is)
  - PowerHouse sensor data
  - Stock data?
- Relate back to filter vs. wrapper methods - give results with different methods?

Chapter 7

Conclusion

7.1 Outline

Not intented for the reader.

- Conclude research
- Future research

Appendix A

Summaries

Please ignore this Appendix. This appendix is for my own personal use. It contains summaries of articles I have read.

A.1 Support Vector Machines

A.1.1 Machine learning: the art and science of algorithms that make sense of data

Book by Peter Flach: [20]. Mainly about chapter 7, "Linear Models". Most important: section 7.3 - 7.5, about support vector machines and non-linearity. **Some parts are direct text; do not use this text directly!**

A.1.1.1 Linear models

Models can be represented by their geometry of d real-values features. Data points are represented in the d-dimensional Cartesian coordinate system/space $\mathcal{X} = \mathbb{R}^d$. Geometric concepts such as lines and planes can be used for *classification* and *regression*. An alternative approach is to use the distance between data points as a similarity measure, resulting from the geometrical representation. Linear methods do not use that property, but rely on understanding of models in terms of lines and planes.

Linear models are of great interest in machine learning because of their simplicity. A few manifestations of this simplicity are:

- Linear models are *parametric*, thus fixed small number of parameters that need to be learned from the data.

- Linear models are *stable*, thus small variations in training data have small impact on the learned model. In logical models they can have large impact, because "splitting rules" in root have great impact.
- Due to relative few parameters, less likely to *overfit* the training data.

The last two are summarized by saying that *linear models have low variance but high bias*. This is preferred with limited data and overfitting is to be avoided.

Linear models are well studied, in particular for the problem of linear regression. This can be solved by the *least-squares* method and classification as discussed in section A.1.1.2, the *perceptron* as explained in section A.1.1.3. Linear regression with the *support vector machine* is handled in section A.1.1.4 and used for probability density estimation in section A.1.1.5. The kernel trick used for learning non-linear models is explained in section A.1.1.6.

A.1.1.2 Least-squares method

The regression problem is to learn a function estimator $\hat{f} : \mathcal{X} \to \mathbb{R}$ from the examples $(x_i, f(x_i))$ where we assume $\mathcal{X} = \mathbb{R}^d$. The difference between the actual and estimated function values are called *residuals* $\epsilon_i = f(x_i) - \hat{f}(x_i)$. The *least-squares method* finds the estimation $\hat{f}$ by minimizing $\sum_{i=1}^{n} \epsilon_i^2$. Univariate regressesion assumes a linear equation $y = a + bx$, with parameters a and b chosen such that the sum of squared residuals $\sum_{i=1}^{n}(y_i - (a + bx_i))^2$ is minimized. Here the estimated parameter $\hat{a}$ is called the *intercept* such that it goes through the (estimated) pooint $(\hat{x}, \hat{y})$ and $\hat{b}$is the *slope* which can be expressed by the (co)variances: $\hat{b} = \frac{\sigma_{xy}}{\sigma_{xx}}$. In order to find the parameters, take the partial derivatives, set them to 0 and solve for a and b.

Although least-squares is sensitive to outliers, it works very well for such a simple method. This can be explained as follows. We can assume the underlying function is indeed linear but contaminated with random noise. That means that our examples are actually $(x_i, f(x_i) + \epsilon_i)$ and $f(x) = ax + b$. If we know a and b we can calculate what the residuals are, and by knowing σ^2 we can estimate *the probability of observing the residuals*. But since we don't know a and b we have to estimate them, by estimating the values for a and b that maximizes the probability of the residuals. This is the *maximum-likelihood estimate* (chapter 9 in the book).

The least-squares method can be used for a (binary) classifier, by encoding the target variable y as classes by real numbers -1 (negative) and 1 (positive). It follows that $\boldsymbol{X}^T(y) = P\boldsymbol{\mu}^+ - N\boldsymbol{\mu}^-$, where P, N, $\boldsymbol{\mu}^+$ and $\boldsymbol{\mu}^-$ are the number of positive and negative

examples, and the d-vectors containing each feature's mean values, resp. The regression equation $y = \bar{y} + \hat{b}(x - \bar{x})$ can be used to obtain a decision boundary. We need to determine the point (x_0, y_0) such that y_0 is half-way between y^+ and y^- (the positive and negative examples, i.e. $y_0 = 0$).

A.1.1.3 Perceptron

Labeled data is *linearly separable* if the exists a linear boundary separating the classes. The least-squares may find one, but it is not guaranteed. Image a perfect linearly separable data set. Move all the positive points away from the negative, but one. At one point the new boundary will exclude (mis qualify) the one original positive outlier, due to the mean-statistics it relies on. The *perceptron* will guaranteed perform perfect separation when the data allows it to be. It was originally proposed as a *simple neural network*. It works by iterating over the training set and modifying the weight vector for every misclassified example ($\boldsymbol{w} \cdot \boldsymbol{x}_i < t$ for positive examples $\boldsymbol{x}_i$). It uses a learning rate η, for a misclassified $y_i = \{-1, +1\}$: $\boldsymbol{w}' = \boldsymbol{w} + \eta y_i \boldsymbol{x}_i$. The algorithm can be made *online* by processing a stream of data points and and updating the weight vector only when a new data point is misclassified.

When the algorithm is completed, every $y_i \boldsymbol{x}_i$ is added α_i times to the weight vector (every time it was misclassified). Thus, the weight vector can be expressed as: $\boldsymbol{w} = \sum_{i=i}^{n} \alpha_i y_i \boldsymbol{x}_i$. In other words: the weight vector is a linear combination of the training instances. The dual form of the algorithm learns the instance weights α_i rather than the features weights $\boldsymbol{w}_i$. An instance $\boldsymbol{x}$ is then classified as $\hat{y} = sign(\sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \cdot \boldsymbol{x})$. This means that during the training only the pairwise dot-products of the data is needed; this results in the n-by-n Gram matrix $\boldsymbol{G} = \boldsymbol{X}\boldsymbol{X}^T$. This instance-based perspective will be further discussed in section A.1.1.4 about the support vector machine.

A.1.1.4 Support Vector Machine

A training example can be expressed by its *margin*: $c(x)\hat{s}(x)$, where $c(x)$ is $+1$ for positive and -1 for negative examples and $\hat{s}(x)$ is the score. The score can be expressed as $\hat{s}(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} - t$. A true positive example $\boldsymbol{x}_i$ has a margin $\boldsymbol{w} \cdot \boldsymbol{x}_i > 0$ and a true negative $\boldsymbol{x}_j$ has $\boldsymbol{w} \cdot \boldsymbol{x}_j < 0$. If m^+ and m^- are the smallest positive and negative examples, then we want the sum of these to be as large as possible. *The training examples with these minimal values are closest to the decision boundary t and are called the support vectors.* The decision boundary is defined as a linear combination of the support vectors. The margin is thus defined as $\frac{m}{\|\boldsymbol{w}\|}$. Minimizing the margin (which is often set to 1 and rescaling is allowed) yields to minimizing $\|\boldsymbol{w}\|$, or: $\frac{1}{2}\|\boldsymbol{w}\|^2$, restricted that none of the

training points fall inside the margin. This gives the following quadratic, constrained optimization problem:

$$\boldsymbol{w}^*, t^* \in \underset{\boldsymbol{w},t}{\operatorname{argmin}} = \frac{1}{2}\|w\|^2 \quad \text{subject to } y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i - t) \geq 1, 1 \leq i \leq n \tag{A.1}$$

This equation can be transformed with the Lagrange multipliers by adding the constraints to the minimization part with multipliers α_i. Taking the partial derivative with respect to t and setting it to 0, we find that for the optimal solution (threshold) t we have $\sum_{i=1}^{n} \alpha_i y_i = 0$. When we take the partial derivative with respect to $w\boldsymbol{w}$ we see that the Lagrange multipliers define the weight vector as a linear combination of the training examples. This partial derivative is 0 for an optimal weight we get that $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i$, *which is the same expression as for the perceptron derived in section A.1.1.3*. By plugging $\boldsymbol{w}$ and t back into the Lagrange equation, we can eliminate these and get the dual optimization problem entirely formulated in terms of the Lagrange multipliers:

$$\Lambda(\alpha_1, \ldots, \alpha_n) = -\frac{1}{2} \sum_{i=i}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j + \sum_{i=1}^{n} \alpha_i \tag{A.2}$$

The dual problem maximizes this function under positivity constraints and one equality constraint: *** TODO: fix equation (now commented) ***

$$\text{subject to } \alpha_i \geq 0, 1 \leq i \leq n \text{ and } \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{A.3}$$

This shows to important properties:

1. Searching for the maximum-margin decision boundary is equivalent to searching for the support vectors; they are the training examples with non-zero Lagrange multipliers.
2. The optimization problem is entirely defined by pairwise dot products between training instances: the entries of the Gram matrix.

The second property enables powerful adaption for support vector machines to learn non-linear decision boundaries, as discussed in section A.1.1.6.

An other solution to non-linear separable data, that is when the constraints $\boldsymbol{w} \cdot \boldsymbol{x}_i - t \geq 1$ are not jointly satisfiable, is to add *slack variables* ξ_i, one for each example. This allows them to be in the margin, of even at the wrong side of the boundary – known as boundary errors. Thus, the constraints become $\boldsymbol{w} \cdot \boldsymbol{x}_i - t \geq 1 - \xi_i$.

"In summary, *support vector machines are linear classifiers that construct the unique decision boundary that maximizes the distance to the nearest training examples (the support vectors).* Training an SVM involves solving a large quadratic optimization problem and is usually best left to a dedicated numerical solver."

A.1.1.5 Density Functions from linear classifiers

The score of an data point can be used to obtain the signed distance of $\boldsymbol{x}_i$ to the decision boundary:

$$d(\boldsymbol{x}_i) = \frac{\hat{s}(\boldsymbol{x}_i)}{\|w\|} = \frac{\boldsymbol{w} \cdot \boldsymbol{x}_i - t}{\|w\|} = \boldsymbol{w}' \cdot \boldsymbol{x}_i - t' \tag{A.4}$$

where $\boldsymbol{w}' = \boldsymbol{w}/\|\boldsymbol{w}\|$ rescaled to unit length and $t' = t/\|w\|$ corresponds to the rescaled intercept. this geometric interpretation of the scores enables them to turn into probabilities. Let $\bar{d}^+ = \boldsymbol{w} \cdot \boldsymbol{\mu}^+ - t$ denote the mean distance of the positive examples to the boundary, where $\boldsymbol{\mu}^+$ is the mean of positive examples (in the grid) and $\boldsymbol{w}$ is unit length. We can assume that the distance of the examples is normally distributed around the mean (which give a bell curve when plotted).

If we obtain a new point $\boldsymbol{x}$ we can get the class by $sign(d(\boldsymbol{x}))$. We would like, instead, to get the probability (using Bayes' rule)

$$\hat{p}(\boldsymbol{x}) = P(+|d(\boldsymbol{x}) = \frac{P(d(\boldsymbol{x})|+)P(+)}{P(d(\boldsymbol{x})|+)P(+) + P(d(\boldsymbol{x})|-)P(-)} = \frac{LR}{LR + 1/clr} \tag{A.5}$$

where LR is the likelihood ratio obtained from the normal score distributions, and clr is the class ratio. With some rewriting we can convert d into a probability by means of the mapping $d \mapsto \frac{exp(d)}{exp(d)+1}$, which is the *logistic function.* The logarithm of the likelihood ratio is linear in $\boldsymbol{x}$ and such models are called *log-linear models.* This logistic calibration procedure can change the location of the decision boundary but not ts direction. There may be an alternative weight vector with a different direction that assign a higher likelihood to the data. Finding that maximum-likelihood linear classifier using the logistic model is called *logistic regression.*

A.1.1.6 Non-linear models

Linear methods such as least-squares for regression can be used for binary classification, yielding in the basic linear classifier. The (heuristic) perceptron guaranteers to classify correctly linear separable data points. Support vector machines find the unique decision boundary with maximum margin and can be adapted to non-linear separable data. These methods can be adjusted to learn non-linear boundaries. The main idea is to

transform the data from the *input space* non-linearly to a *feature space* (which can, but does need to be in a higher dimension) in which linear classification can be applied. The mapping back from the feature space to the input space is often non-trivial (e.g. mapping (x, y) to feature space by (x^2, y^2), yields in four coordinates when transformed back to the input space).

The remarkable thing is that often the feature space does not have to be explicitly constructed, as we can perform all necessary operations in input space. For instance; the perceptron algorithm mainly depends on the dot product of $\boldsymbol{x}_i \cdot \boldsymbol{x}_j$. Assuming $\boldsymbol{x}_i = (x_i, y_i)$ and $\boldsymbol{x}_j = (x_j, y_j)$, the dot product can be written as $\boldsymbol{x}_i \cdot \boldsymbol{x}_j = x_i x_j + y_i y_j$. The instances in quadratic feature space are (x_i^2, y_i^2) and (x_j^2, y_j^2) and their dot product is $(x_i^2, y_i^2) \cdot (x_j^2, y_j^2) = x_i^2 x_j^2 + y_i^2 y_j^2$. This is almost equal to $(\boldsymbol{x}_i \cdot \boldsymbol{x}_j)^2 = (x_i x_j)^2 + (y_i y_j^2 + 2x_i x_j y_i y_j$, but not quite because of the third term. We can make the equations equal by *extending the feature space* (to a higher dimension) with a third feature $\sqrt{2xy}$, so the feature space is $\phi(\boldsymbol{x}_i) = (x_i^2, y_i^2, \sqrt{2x_i y_i})$.

If we define $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i, \boldsymbol{x}_j)^2$ and replace $\boldsymbol{x}_i \cdot \boldsymbol{x}_j$ with $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ in the (perceptron) algorithm, we obtain the *kernel perceptron* with the degree $p = 2$. We are not restricted to polynomial kernels; an often used kernel is the *Gaussian kernel*, defined as:

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}) \tag{A.6}$$

where σ is known as the *bandwidth* parameter. We can think of the Gaussian kernel as imposing a Gaussian (i.e. , multivariate normal) surface on each support vector in instance space, so that the boundary is defined in therms of those Gaussian surfaces. Kernel methods are best known in combination with support vector machines. Notice that the soft margin optimization problem is defined in terms of dot product between training examples, and thus the 'kernel trick' can be applied. Note that the decision boundary learn with a non-linear kernel cannot be represented by a simple weight vector in input space. Thus, to classify a new example $\boldsymbol{x}$ we need to evaluate $y_i \sum_{j=1}^{n} \alpha_j y_j \kappa((x), \boldsymbol{x}_j)$ (the Gram matrix?) involving all training examples, or at least all with non-zero multipliers a_j (the support vectors).

A.1.2 Change Point Detection In Time Series Data Using Support Vectors

Paper by Fatih Camci [13] About segmentation with SVMs. Will be main material for section 2.6 about SVMs.

A.1.2.1 Introduction

Interprets change detection as finding the transition points from one underlying time series generation model to another. The change point is mostly represented in a sudden change in mean or variance. Existing models detect changes in mean and increase in variance, *but fail to recognize decrease in variance.* Many methods require some model (like Auto-Regressive [AR]) to fit the time series in order to eliminate the noise. Thus, the effectiveness of the method is tied to the fitness degree of the model to the time series data. These two problems (lack of variance decrease detection and model-bound fitness degree) leads to this work; Support Vector based Change Point Detection targeting changes in variance and/or mean without any assumption of model fitting of data distribution. This method *does not use a time series model* for fitting and targets *both increase and decrease* in mean and variance.

A.1.2.2 Related work

Change Point Detection (CPD) can be categorized in posterior (off-line) and sequential (on-line). Sequential receive data sequentially and analyze previously obtained data to detect the possible change in current time. This method is based on sequential analysis and focuses on change on mean and variance in time domain. Other methods generally suffer from:

- Inability / inefficiency in detecting variance decrease.
- Assumptions about the statistical distribution of the data, obtained as error of fitting the (AR) model.
- Necessity of training the model with possible changes.

A.1.2.3 Support vector based one-class classification

Although SVM was originally designed for two-class classification, it has been successfully applied to multi-class and one-class classification. SVM-based one-class classification gives the minimum volume closed spherical boundary around the data, represented by center c and radius r. It minimizes r^2 (representing structural error), and uses a penalty coefficient C for each outlier with distance ξ_i from the hyper-sphere boundary:

$$\begin{gathered} \text{Min } r^2 + C\sum_i \xi_i \\ \text{Subject to : } \|\boldsymbol{x}_i - c\|^2 \leq r^2 + \xi_i \quad \xi_i \geq 0, \quad \forall i, \boldsymbol{x}_i : i\text{th data point} \end{gathered} \tag{A.7}$$

This quadratic optimization problem can be transformed to its dual form by introduction Lagrange multipliers α_i. If, for a data point, the multiplier $\alpha_i = 0$, then that point is inside the sphere. When it is $0 < \alpha_i < C$, then it is on the boundary. Data points for which the multiplier is $\alpha_i = C$ are located outside the sphere (and are penalized). The dual form is:

$$\begin{aligned} \text{Max} \sum_i \alpha_i(\boldsymbol{x}_i \cdot \boldsymbol{x}_i) - \sum_{i,j} \alpha_i \alpha_j (\boldsymbol{x}_i \cdot \boldsymbol{x}_j) \\ \text{Subject to : } 0 \leq \alpha_i \leq C \quad \forall i, \quad \sum_i \alpha_i = 1 \end{aligned} \tag{A.8}$$

Note that only dot-products of the data points $\boldsymbol{x}$ appear. In order to transform the data points to a higher dimension, to create a good representational hyper-sphere, kernels replace the dot products without compromising computational complexity. The problem then becomes:

$$\text{Max} \sum_i \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}_i) - \sum_{i,j} \alpha_i \alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{A.9}$$

It has been shown that Gaussian kernels offer better performance for one-class classification the others. The optimization of the *scale parameter* has led to several implementations. As can been seen, there are no assumptions about the data distribution or independency made.

A.1.2.4 Problem formulation

[Not summarized here, useful for e.g. section 2.2]

A.1.2.5 SVCPD: The algorithm

Instead of using statistical properties of the data, for each window of size w a hyper-sphere is constructed without increasing computational complexity due to the *kernel trick*. The window size is related to the sensitivity of the method to change; small windows are sensitive with high false alarm rate whilst large windows are slow to detect change and have low alarm rates. The algorithm is listed in table A.1. Note that SVCPD can be applied directly to multidimensional data, whilst many other methods can only be applied to one-dimensional data.

Step	Action
1	Start with n observations and construct hyper-sphere
2	Add next observation x_t and drop first one
3	Identify new hyper-sphere and its *approximate radius*
4	if x_t is outside hyper-sphere, mark t as change points and continue from step 2
5	calculate radius average of last w hyper-planes
6	calculate radius ratio $\hbar$. If lower than th_{low} or greater than th_{high} then mark t as change point
7	continue from step 2

TABLE A.1: Support Vector machine based Change Point Detection algorithm

A.2 CUSUM for variance

A.2.1 Use of Cumulative Sums of Squares for Retrospective Detection of Changes of Variance

Carla Inclan and George C. Tiao [33], 1944, 162 refs.

A.2.1.1 Introduction

This paper is about reflective detection of multiple changes of variance in a sequence of independent observations. This is a statistical method, which differs from others (in that field) such as Bayesian method (Bayes ratio, posterior odds), maximum likelihood methods and (autoregressive) models. This approach uses the centered version of cumulative sums of squares to search for change points systematically and iteratively (and reflective).

A.2.1.2 Centered Cumulative Sum of Squares

The cumulative sum of squares is often used for change detection in the mean. It is defined as $C_k = \sum_{i=1}^{k} \alpha_t^2$ for a series of uncorrelated random variables $\{\alpha_t\}$ with mean 0 and variance $\sigma_t^2, t = 1, 2, \ldots, T$. The centered (and normalized) cumulative sum of squares is:

$$D_k = \frac{C_k}{C_T} - \frac{k}{T}, \quad k = 1, \ldots, T, \quad \text{with } D_0 = D_T = 0 \tag{A.10}$$

For homogeneous variance the plot of D_k against k (the first k elements of the series) will oscillate around 0. When a sudden change in variance occurs, the pattern of the plot of D_k will break out some specified boundaries with high probability. For C_k it holds that, under homogeneous variance, the plot will be a straight line with slope σ^2.

With one of more change points the plot appears as a line of several straight pieces. The plot of D_k creates a peak for a smaller and a trough for a larger variance, is visually more clear and breaks out a predefined value. The search for a change point is variance is than to find $k^* = \max_k |D_k|$. If the value of D_k at k^* exceeds a predefined value (e.g. $D^*_{0.5} = 1.358$, for $\sqrt{T/2D_k}$ because of the Brownian bridge property), that value of k^* will be an estimate for a change point.

There is a relation between D_k and the F statistic, which is used for testing equality of variances between two independent samples. For a fixed k, $D_k(F)$ is a monotone function of F (it depends only on k through k/T). An important distinction: the F statistic is *used with known* k, whereas we are looking for $\max_k |D_k|$ to determine the location of the change point.

When we assume that $\{\alpha_t\}$ is normally distributed with mean 0 and variances σ_t^2, then we can obtain the *likelihood ratio* for testing the hypothesis of one change against the hypothesis of no change in the variance. When maximizing the likelihood estimator for a location κ, we can find the log-likelihood ratio $LR_{0,1}$. Although $LR_{0,1}$ and $\max_k |D_k|$ are related, they are not the same. The latter puts more weight near the middle of the series is thus biased toward $T/2$.

The (expected) value of D_k given a change in variance differs in the context. If a smaller variance corresponds to the smaller portion of the series, then it will be harden to find the change point using D_k. There is a masking effect when there are multiple change points in the series; the order of small, medium and large variances result in the value of D_k. The iterative algorithm presented in this paper in section A.2.1.3 is designed to lessen the masking effect.

A.2.1.3 Multiple changes: Iterated cumulative sums of squares

In case of a single change point the D_k method would succeed. But we are interested in multiple change points of variance, and thus the usefulness of the D_k reduces due to the masking effect. A solution is to iteratively applying the method and dividing the series at each possible change point. The algorithm is presented in table A.2. It is the third steps which reduces the masking effect and helps to "fine tune" the algorithm by (re)moving the potential change points by checking each location given the adjacent ones.

Step	Action
0	Let $t_1 = 1$
1	Calculate $D_k(\alpha[t_1 : T])$. Let $k^*(\alpha[t_1 : T])$ be the point at which $\max_k \lvert D_k(\alpha[t_1 : T])\rvert$. Let D^* be the asymptotically critical value and M the max value in the series segment (?). If $M > D^*$ then consider k^* to be a change point. Else, there is no change point and the algorithm stops.
2a	Repeat for the first part (up to the change point), until no more change points are found.
2b	Repeat for the second part (from the change point forward), until no more change points are found.
3	When two or more change points are found; check for each $\alpha[j-1 : j+1]$ if there is indeed a change point (j). Repeat until the number of change points does not change and each new found change point is "close" enough to previous.

TABLE A.2: Iterated Cumulative Sums of Squares Algorithm

A.2.1.4 Results

When the ICSS algorithm was applied to stock data, it resulted in comparable results as the maximum likelihood estimates and Bayesian analysis. The performance (CPU-time and correct observations with artificial data) of ICSS outperforms the other two. The heavy computational burden of posterior odds can be partially alleviated by the maximum log-likelihood method. The ICSS algorithm avoids calculating a function at all possible locations of change points due the iterative manner.

A.3 Density Ratio Estimation

A.3.1 Change-Point Detection in Time-Series Data by Direct Density-Ratio Estimation

Kawahara and Sugiyama [35], 2009, 45 refs.

"This paper provides a change-point detection algorithm based on direct density-ratio estimation that can be computed very efficiently in an online manner".

A.3.1.1 Introduction

The problem of change-point detection is well studied over the last decades in the field of statistics. A common statistical formulation of change-point detection is to consider the probability distributions over past and present data intervals, and regard the target time

as a change-point if the two distributions are significantly different. Some approaches (such as CUSUM and GLR) make use of the *log-likelihood ratio*, and are extensively explored in the data mining community. Many approaches (novelty detection, maximum-likelihood ratio, learning of autoregressive models, subspace identification) rely on pre-specified parametric models (probability density models, autoregressive models, state-space models). That makes it less applicable to real-life problems. There have been some non-parametric density estimation approaches proposed, but that is known to be a hard problem. The key idea of this paper to directly estimate the *ratio* of the probability densities (also known as *importance*). The KLIEP is an example, but it is a batch algorithm. This paper introduces an online version of the KLIEP algorithm and develops a flexible and computationally efficient change-point detection method. This method is equipped with a natural cross validation procedure and thus the value of tuning parameters can be objectively determined.

A.3.1.2 Problem formulation and Basic Approach

Let $\boldsymbol{y}(t)(\in \mathbb{R}^d)$ be a d-dimensional time series sample at time t. The task is to detect whether there exists a change point between two consecutive time intervals, called the *reference* and *test* intervals. The conventional algorithms consider the likelihood ratio over samples from the two intervals. Since time-series samples generally are not independent over time it is hard to deal with them directly. To overcome this difficulty, we consider *sequences* of samples in the intervals: $\boldsymbol{Y}(t)(\in \mathbb{R}^{dk})$ is the forward subsequence of length k at time t. This is a common practice in subspace identification since it takes implicitly time correlation into consideration. The algorithm in the paper is based on the logarithm of the likelihood ration of the *sequence sample* $\boldsymbol{Y}$:

$$s(\boldsymbol{Y}) = \ln \frac{p_{\mathrm{te}}(\boldsymbol{Y})}{p_{\mathrm{rf}}(\boldsymbol{Y})} \tag{A.11}$$

where $p_{\mathrm{te}}(\boldsymbol{Y})$ and $p_{\mathrm{rf}}(\boldsymbol{Y})$ are the probability density functions of the reference and test sequence samples. Let t_{rf} and t_{te} be the starting points of the reference and test intervals. Decide if there is a change-point between the reference and test interval by monitoring the logarithm of the likelihood ratio:

$$S = \sum_{i=1}^{n_{\mathrm{te}}} \ln \frac{p_{\mathrm{te}}(\boldsymbol{Y}_{\mathrm{te}}(i))}{p_{\mathrm{rf}}(\boldsymbol{Y}_{\mathrm{te}}(i))} \tag{A.12}$$

If, for a predefined $\mu > 0$, it holds that $S > \mu$ then a change occurs. The remaining question is how to calculate the density ratio, because it is unknown and we need to estimate it from examples. The naive approach would be to first estimate the densities for the reference and test interval separately and then take the ratio. This approach

via non-parametric density estimation may not be effective — directly estimating the density ratio without estimating the densities would be more promising.

The direct estimation of the density ratio is based on the Kullback-Leibler Importance Estimation Procedure. Let us model the density ratio $w(\boldsymbol{Y})$ by a non-parametric Gaussian kernel model:

$$\hat{w}(\boldsymbol{Y}) = \sum_{l=1}^{n_{\mathrm{te}}} \alpha_l K_\sigma(\boldsymbol{Y}, \boldsymbol{Y}_{\mathrm{te}}(l)), \tag{A.13}$$

where $\{\alpha_l\}_{l=1}^{n_{\mathrm{te}}}$ are parameters to be learned from the data samples and $K_\sigma(\boldsymbol{Y}, \boldsymbol{Y}')$ is the Gaussian kernel function with mean $\boldsymbol{Y}'$ and standard deviation σ.

A.3.1.3 Online Algorithm

...

A.4 Outlier Detection methods

A.4.1 A Survey of Outlier Detection Methodologies

Hodge and Austin, [30], 2004, 734 refs.

A.4.2 Summary

This survey takes a look at different methodologies that perform outlier detection. It gives two definitions of outliers, one of which relates to the problem statement in this thesis (taken from [8]):

> An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.

The survey introduces three fundamental approaches to the problem of outlier detection:

Type 1 - no prior knowledge about the data; analogous to *unsupervised clustering*. Flags the remote points as outliers; mainly batch-processing systems.

Type 2 - model (and requires) both normal and abnormal data; analogous to *supervised classification*. Is able to process new data online.

Type 3 - Model only normal data (and in very few cases abnormal). Generally referred to as *novelty detection*, analogous to *semi-supervised recognition or detection*. It only requires pre-classified normal data. It aims to define a boundary of normality.

The type of approaches of interest in this research is of **type 3**. It is characterized by the ability to recognize new data as normal when it lies within the constructed boundary and as a novelty otherwise. This ability removes the need of sample-abnormality data, which may be hard (or costly) to produce. The method of Tax *et al.* [57] is stated to be of **type 2**, and one can argue that other methods of Tax *et al.* [58, 60] are by their one-class classification instances of **type 3** methods.

The survey states that density-based methods estimate the density distribution of the training data. Outliers are identified as data points lying in a low-density region.

Appendix B

Session with Anne 24-06-2013 - Paper Camci analysis

Please ignore this Appendix. This appendix is for my own personal use. This chapter will look at the paper of Camci [13] ("Change point detection in time series using support vectors") and will answer many question that the paper leaves open. The goal is to make a better justification for the used techniques and made assumptions.

B.1 Density estimation / Data description / Vapnik's principle

Following Vapnik's principle, one should "When solving a problem of interest, do not solve a more general problem as an intermediate step" [61] when a limited amount of data is available. For the problem of change detection we are only interested in some characteristics of the data. Solving the complete density estimation might require more data than actually needed when the requested characteristic is a closed boundary around the data.

In [40] it is stated that the SVM by Cortes and Vapnik [19] is a representative example of this principle. Instead of estimating the more general data generating probability distributions, it only learns a decision boundary to differentiate between the two distributions.

The proposed method SVDD of Tax and Duin [59, 60] models the boundary of data under consideration. Thereby it characterizes a data set and can be used to detect novel data or outliers. The performance is compared to methods which model the

distribution's density instead, using Receiver-Operating Characteristic (ROC) curves and false negative rates. The compared methods are: *(1)* normal density which estimates the mean and covariance matrix; *(2)* the Parzen density where the width of the Parzen kernel is estimated; *(3)* a Gaussian Mixture Model optimized using EM; and *(4)* the Nearest-Neighbor Method which compares the local density of an object with the density of the nearest neighbor in the target set (and is thus, just as SVDD, a boundary-based method). The results show that when the problem formulation is to characterize an area in a feature space (and not the complete density distribution) SVDD gives a good data description.

The study of Tax on One-Class Classifiers (OCCs) [58] further compares density methods, boundary methods (amongst which SVDD) and reconstruction methods. One promesing result for SVDD is that it performs well on read-world data, for which generalization is needed.

The method of Camci [13] uses a Support Vector based method to find change points in the data. It does not explicitly create a density estimation, but instead relies on the spherical boundary and uses its ability to detect novel data or outliers to detect change points in time series data.

**** Thus methods of data descriptions should be compared (which are more general) and not only density estimation? ****

The paper of Yin *et al.* [66] makes a distinction between similarity based (using a defined distance measure) and model based (which characterize the data using predictive models) approach. The OC-SVM is used in the model-approach to filter out normal activities in order to detect abnormality behavior.

B.2 Change point definition

The method of Camci [13] regards a change point as the moment in time that the underlying stochastic process has changed, say from p^1 to p^2. It assumes that each of these stochastic processes is modeled following a Gaussian distribution, such that a change can occur in the value of the mean and/or the variance; $p^1 \sim N(\mu_1, \sigma_1^2)$. The CUSUM-based method of [33] also regards each semgent as a Gaussian distribution.

The method if Kawahara *et al.* [35] is based on the log likelihood ratio of test samples, and the method by Liu *et al.* [40] uses a comparable dissimilarity measure using the KLIEP algorithm.

The method of Chamroukhi *et al.* [15] is based on a Hidden Markov Model and logistic regression. It assumes a K-state hidden process with a (hidden) state sequence, each state providing the parameters (amongst which the order) for a polynomial. The order of the model segment is determined by model selecting, often using the BIC or the similair AIC [3], as in [28].

*** REVIEW zoeken: change points in time series ***

Periodicity/consecutive data vs unique/irregular data?

Change in model parameters (mean/variance, of linear/non-linear)? –¿ Model selection?

Definition of continuity –¿ windows the domain and problem. Will result in definition of dis-continuity –¿ this is the goal to find Relation with double differentiation.

B.3 Data and model

Why assume (model of) data is Gaussian/normal distribution? Thus, piecewise linear with mean and variance as changing properties. Why not set of polynomial models, as for example in [15]?

What is the best model for accelerometer data of human activities? –¿ Look to result of model-selecting papers.

Should we build a model of the data? (Gaussian distribution is also the model). And be able to reconstruct?

Why is it better (as Camci states) that a method makes no assumptions about the form of data/distribution of data? Is it that there are less parameters to estimate?

Many methods describe and compare methods to construct classifier models for the classification of accelerometer data, such as [38] and [67] (often using extracted features from the raw data signal). In contrast, we could not find a clear characterization of accelerometer data obtained from human activities. When the problem of temporal segmentation is regarded in this context, a formalization of the data under consideration is needed. Some assume the data follows a piecewise linear set of segments with a mean and noise/variance modeled by a normal (Gaussian) distribution (such as [13]). Other approaches regard the data as a set of polynomials, which can be estimated by regression (such as [15]) and apply a form of model selection to each segment.

*** TODO: make a clear distinction between model types; similarity (distance based) or model based, as in [66] ***

B.4 Segmentation (SVM) method

Overview of segmentation methods. Collection of papers use relative and direct density-ratio estimation [35, 40].

Why use SVM for density estimation? Look for justification, perhaps a review paper which compares/mentions SVM for temporal segmentation of (human activity type of) data?

Why use RBF/Gaussian kernel? Is it because of OC-SVM of because of form the data? Why not polynomial/linear?

- "Use RBF when relation between class and data is non-linear"
- "RBF uses less paramters (C for penalty/soft margin and gamme for kernel width) than non-linear polynomial kernels"
- (arguments from [67] "Optimal model selection for posture recognition in home-based healthcare")
- Survey [30] states RBF is similair to Gaussian Mixture model (page 25).

*** TODO: read [45] and [10] on geometry of SVMs ***

Re-evaluatie [24], uses (amongst others) a pca-based dimension-detection method (?). Also uses model-selection as an intermediate step.

B.4.1 Higher dimension mapping (including kernel)

What does the mapping from the data-space to higher dimension looks like?

What is a RBF kernel?

What form has the higher dimensional space?

How do relations over data, such as distance, volume and noise, act in the higher dimensional space?

What is the kernel trick to not explicit do the mapping to the higher space?

*** Note: in [45] it is explained why the inner-product between two vectors is a logical choice for the distance/similarity measure. ***

B.5 Relation to other methods

B.5.1 Novelty/outlier detection

*** Read [30] "A survey of outlier detection methodologies" to compare with other methodologies. ***

- One-class classification is referred to as "Type 3", with *semi-supervised recognition or detection.*
- It explains why one-class can be beneficial over type-2 where negative examples needs to be provided.
- We are interested in type-3, so compare SVM with the others stated in this survey regarding type-3.
- Often refers to the convex hull of the data set. Link with geometric approach as in [10, 45]?
- It states that PCA and regression techniques are linear models and thus often are too simple for practical applications. SVMs try to find a hyperplane in higher dimensional space; linear models to implement complex class boundaries. It refers to [59].

B.5.2 Scale parameter

B.5.3 Robust statistics / "M-Estimators"

Is the method robust, in the sense that outliers have restricted impact on the quality.

What is the relation to M-Estimators (Wikipedia: "M-estimators are a broad class of estimators, which are obtained as the minima of sums of functions of the data", http://en.wikipedia.org/wiki/M-estimator)

B.6 Quality metrics

As used in Camci: benefit, false alarm rate

Asymmetric test: feed data from front-to-back and back-to-front; how far are matched datapoint apart.

Model reconstruction error: try to reconstruct the simulated models, test similarity (BIC?). Is that a good measure? (If we are only interested in finding change points...)

Appendix C

Summary of papers and principle formulas

C.1 Change point detection in time series data using support vectors, by Camci

Paper: [13]. The main concept of this paper is to construct a hypersphere around the data and thereby generating a boundary. A change point is detected when the radius of the hypersphere grows or shrinks significantly, or when a data point falls outside the boundary.

The main cost function being minimized:

$$\begin{aligned} &\underset{r}{\text{minimize}} \quad r^2 + C\sum_i \xi_i \|x\| \\ &\text{subject to} \quad \|\boldsymbol{x}_i - c\|^2 \leq r^2 + \xi_i, \quad \xi_i \geq 0 \quad \forall i, \boldsymbol{x}_i : i\text{th data point} \end{aligned} \tag{C.1}$$

Where r is the radius of a (hyper)circle with center $\boldsymbol{c}$, C is the penalty coefficient for every outlier and ξ_i is the distance from the ith data point to hypersphere (also known as the slack variable).

The dual form by introducing the Lagrange multipliers ($\alpha_i, \alpha_i \geq 0$) and eliminating the slack variables ξ is:

$$\begin{aligned} &\underset{\boldsymbol{\alpha}}{\text{maximize}} \quad \sum_i \alpha_i(\boldsymbol{x}_i \cdot \boldsymbol{x}_i) - \sum_{i,j} \alpha_i \alpha_j (\boldsymbol{x}_i \cdot \boldsymbol{x}_j) \\ &\text{subject to} \quad 0 \leq \alpha_i \leq C \quad \forall i, \quad \sum_i \alpha_i = 1 \end{aligned} \tag{C.2}$$

To allow for a non-linear relation between the data points and the data boundary, the inner product can be replaced by a (e.g. Gaussian) kernel function: $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

C.2 Change-Point detection in time-series data by Direct Density-Ratio Estimation

Paper: [35].

The density ratio $w(\boldsymbol{Y})$ is modeled by a Gaussian kernel over sequences $\boldsymbol{Y}$ of samples (sequence $\boldsymbol{Y}_{te}(l)$ is the test sequence from the lth position on):

$$\hat{w}(\boldsymbol{Y}) = \sum_{l=1}^{n_{\mathrm{te}}} \alpha_l K_\sigma(\boldsymbol{Y}, \boldsymbol{Y}_{\mathrm{te}}(l)), \tag{C.3}$$

where $\{\alpha_l\}_{l=1}^{n_{\mathrm{te}}}$ are parameters to be learned from the data samples and $K_\sigma(\boldsymbol{Y}, \boldsymbol{Y}')$ is the Gaussian kernel function with mean $\boldsymbol{Y}'$ and standard deviation σ. The learned parameters minimize the Kullback-Leibler divergence from the sequence to the test sequence. The maximization problem then becomes:

$$\begin{aligned} \underset{\{\alpha_l\}_{l=1}^{n_{\mathrm{te}}}}{\text{maximize}} \quad & \sum_{i=1}^{n_{\mathrm{te}}} \log \left(\sum_{l=1}^{n_{\mathrm{te}}} \alpha_l K_\sigma(\boldsymbol{Y}_{\mathrm{te}}(i), \boldsymbol{Y}_{\mathrm{te}}(l)) \right) \\ \text{subject to} \quad & \frac{1}{n_{\mathrm{rf}}} \sum_{i=1}^{n_{\mathrm{rf}}} \sum_{l=1}^{n_{\mathrm{te}}} \alpha_l K_\sigma(\boldsymbol{Y}_{\mathrm{rf}}(i), \boldsymbol{Y}_{\mathrm{te}}(l)) = 1, \text{ and } \alpha_1, \ldots, \alpha_{n_{\mathrm{te}}} \geq 1 \end{aligned} \tag{C.4}$$

With the estimated parameters the logarithm of the likelihood ratio between the test and reference interval can be calculated, which signals a change point if it is beyond a certain threshold μ:

$$S = \sum_{i=1}^{n_{\mathrm{te}}} \ln \frac{p_{\mathrm{te}}(\boldsymbol{Y}_{\mathrm{te}}(i))}{p_{\mathrm{rf}}(\boldsymbol{Y}_{\mathrm{te}}(i))} \tag{C.5}$$

C.3 Joint segmentation of multivariate time series with hidden process regression for human activity recognition, by Chamroukhi

Paper: [15]. This approach models the time series data by a parameterized regression, filtered with a HMM to smooth out high frequency activity transitions. With each observation i, generated by a K-state hidden process, an activity label z_i (and thus sequence) is associated. Observations follow a regression model:

$$y_i = \beta_{z_i}^T \boldsymbol{t}_i + \sigma_{z_i} \epsilon_i; \quad \epsilon_i \sim \mathcal{N}(0,1), \quad (i = 1, \ldots, n) \tag{C.6}$$

The observations get a label assigned by maximizing the logistic probability (π_k):

$$\hat{z}_i = \underset{1 \leq k \leq K}{\operatorname{argmax}} \, \pi_k(t_i; \hat{\boldsymbol{w}}), \quad (i = 1, \ldots, n) \tag{C.7}$$

C.4 Support Vector Data Description, by Tax and Duin

Paper: [60]. The paper proposes the SVDD method, analogous to the Support Vector Classifier (SVC) of Vapnik [61], based on the separating hyper-plane of Schölkopf *et al.* [49]. Where SVC is ablo to distinguish data between two classes, SVDD obtains a closed boundary around the target class and can detect outliers.

Method and formulas very similair to description in Section C.1.

C.5 Support Vector Density Estimation, by Weston et al.

Paper: [63]. Using the notation of [63], the distribution function of a density function $p(x)$ is represented as:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} p(t)\,\mathrm{d}t \tag{C.8}$$

To find the density the following linear equation need to be solved:

$$\int_{-\infty}^{\infty} \theta(x-t)p(t)\,\mathrm{d}t = F(x) \tag{C.9}$$

where

$$\theta(x) = \begin{cases} 1, & x > 0 \\ 0, & otherwise \end{cases}$$

In this problem the distribution function $F(x)$ is unknown and instead we are given the independently and identically distributed (i.d.d.) data $x_1, \ldots, x_l$ generated by F.

The empirical distribution function can now be constructed as:

$$F_l(x) = \frac{1}{l}\sum_{i=1}^{l} \theta(x - x_i) \tag{C.10}$$

C.6 An online algorithm for segmenting time series, by Keogh et al.

Paper: [37]. This method approximates the signal with piecewise linear representation. Change points are encountered at the time at which a new segment is used to represent the signal. The method uses linear regression by taking the best fitting line in the least squares sense, since that minimizes the Euclidian distance which is used as a quality metric. Linear interpolation is considered but since that always has a greater sum of squares error it is disregarded.

Linear regression assumes a relation from n observations $\boldsymbol{X}$ to the dependend variable $\boldsymbol{y}$ using the parameter vector $\boldsymbol{\beta}$:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon \tag{C.11}$$

The error function, the sum of squared residuals, being minimized by searching for the best estimation of $\boldsymbol{\beta}$ is:

$$b^* \in \underset{b}{\operatorname{argmin}} = \sum_{i=1}^{n} (y_i - x_i' b)^2 = (\boldsymbol{y} - \boldsymbol{X}b)^T (\boldsymbol{y} - \boldsymbol{X}b) \tag{C.12}$$

C.7 Online novelty detection on temporal sequences, by Ma and Perkins

Paper: [42]. This method uses support vectors for regression (in contrast with of classification). The regression function, using a kernel function $K(x_i, x_j)$ can be written as:

$$f(\boldsymbol{x}) = \sum_{i=1}^{l} \theta_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b, \tag{C.13}$$

where θ_i is a coefficient resulting from the Lagrange multipliers of the original minimization problem. A small fraction these of coefficients are non-zero, and the corresponding samples $\boldsymbol{x}_i$ are the *support vectors*. The regression function $f(\boldsymbol{x})$ is non-linear when a non-linear kernel is chosen.

The regression function is used to created a model of past observations. A matching function is constructed which determines the matching value $V(t_0)$ of a new observations with the constructed model. This matching value is the residual of the regression function at t_0.

The algorithm determines *(novel) events*, *occurrences* and *surprises*. *Novel events* are defined as a series of observations for which the confidence value over the number of suprises (out-of-model observations) is high enough. Events thus have a length; they are constructed of a sub-series of observations.

The papers presents an alternative implementation in order to handle fixed-resource environments and thus induce an online algorithm. After W observations have been observed and used for the trained model, the oldest observation is disregarded before the newly obtained observation is incorporated.

Note: the Support Vector approach in this paper is used to select the observations to use in the regression model. This differs from one-class applications of support vector machines. The same authors have also presented a paper which does use one-class construction using support vector machines: [43].

C.8 Time-series novelty detection using one-class support vector machines, by Ma and Perkins

Paper: [43] This approach is very similar to the other paper of this author discussed in the preview section [42]. The difference is that this method does create a SVM-classifier to detect in and out of class examples, whilst the other uses the support vectors to construct a regression function.

The method constructs a hyper-plane which separates as many as possible data points in the feature space with the largest margin from the origin. This is a different from (and more like the original SVM-proposal by Schölkopf [50]) the one-class methodology by Tax which creates a boundary around the data [60].

The hyper-plane in feature space is represented as:

$$f(\boldsymbol{x}) = \boldsymbol{W}^T \Phi(\boldsymbol{x}) - \rho, \tag{C.14}$$

where $\Phi(\boldsymbol{x})$ maps a vector $\boldsymbol{x}$ from the input space I to the (potentially infinite dimensional) feature space F. $\boldsymbol{W}$ and ρ are determined by solving a quadratic optimization problem. The dual formulation (using Lagrange multipliers α_i) is:

$$\boldsymbol{W} = \sum_{i=1}^{l} \alpha_i \Phi(\boldsymbol{x}_i), \tag{C.15}$$

where $0 \leq \alpha_i \leq \frac{1}{\nu l}$. The parameter $\nu \in (0, 1)$ is set to trade-off the smoothness of $f(\boldsymbol{x})$ and acts as a upper bound on the fraction of outliers over all the data examples in $\boldsymbol{x}$ [50].

Using the *kernel trick* the inner product of two vectors in feature space F can be replaced by a kernel function K, which is often the RBF. The equation of the hyper-plane (C.14) then becomes the following (non-linear) function:

$$f(\boldsymbol{x}) = \sum_{i=1}^{l} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) - \rho \tag{C.16}$$

The form of the input vector $\boldsymbol{x}$ is considered to be the *(projected) phase space* representation of the original time series. Just like [36] which constructs *sequences* of samples and in contrast with [13], each element of $\boldsymbol{x}$ is a vector with the size of the *embedding dimension* E of the time series. Thus, a time series $x(t)$ is converted to a set of vectors $T_E(N)$:

$$T_E(N) = \{\boldsymbol{x}_E(t),\ t = E \cdots N\}, \tag{C.17}$$

where

$$\boldsymbol{x}_E(t) = [x(t-E+1)\ x(t-E+2)\ \cdots\ x(t)] \tag{C.18}$$

If a point of this vector $\boldsymbol{x}_E(t)$ is regarded (in the feature space F) as an outlier, all corresponding values in the original time series are also regarded as such.

C.9 Least squares one-class support vector machine, by Choi

Paper: [18]. The proposed method uses a SVM for a similarity/distance comparison for testing examples to training examples. Instead of other methods, such as the standard one-class SVM by Schölkopf [50] or the SVDD of Tax and Duin [60], it does not create a boundary for the training data. Instead it uses a least-squared approach to construct a hyperplane to which most of the training examples lie close to.

The objective function to be minimized of the standard one-class method by Schölkopf is formulated as:

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 - \rho + C\sum_j \xi_j \\ \text{subject to} \quad & \boldsymbol{w}\cdot\phi(\boldsymbol{x}_j) \geq \rho - \xi_j \quad \text{and} \quad \xi_j \geq 0 \end{aligned} \tag{C.19}$$

where ϕ is a mapping to the feature space.

The least-squares one-class support vector machine has a small variation on the above formula, which results in the following minimization problem:

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 - \rho + \frac{1}{2}C\sum_j \xi_j^2 \\ \text{subject to} \quad & \boldsymbol{w}\cdot\phi(\boldsymbol{x}_j) = \rho - \xi_j \end{aligned} \tag{C.20}$$

The slack variable (for which in the original formulation $\xi_j \geq 0$ should hold) now represents an error caused by a training example $\boldsymbol{x}_j$ with relation to the hyperplane, i.e. $\xi_j = \rho - \boldsymbol{w}\cdot\phi(\boldsymbol{x}_j)$.

In other words, the minimization problem (C.20)) results in a hyperplane with maximal distance from the origin and for which the sum of the squares of errors ξ_j^2 are minimized.

Appendix D

Planning

Chapter	Title	Extra	Deadline
1	Introduction, abstract	Write at last	15/10/2013
2	Literature review	Partially writting. Maybe focus more on SVM and quick look at other segmentation techniques	15/9/2013
3	(Change detection By) One-Class Support Vector Machines	Mix of current chapter and blog post	1/10/2013
4	Proposed method	Propose new method, or just setup of testing the OC-SVMs with accelerometer data	Date
5	**Results**	Test with known and public databases of accelerometer data. For this, two implementations (in Matlab?) are needed	15/8/2013
6	Real-world applications	Test with own labeled accelerometer data	1/9/2013
7	Conclusion	Write last, together with introduction	15/10/2013

Appendix E

Data structure quotes

[47], "Predicting Time Series with Support Vector Machines", p.6.
"Our experiments show that SVR methods work particularly well if the data is sparse (i.e. we have little data in a high dimensional space). This is due to their good inherent regularization properties."

[46], "Support vector machines in remote sensing: A review", p.10.
"Most of the findings show that there is empirical evidence to support the theoretical formulation and motivation behind SVMs. The most important characteristic is SVMs ability to generalize well from a limited amount and/or quality of training data. Compared to alternative methods such as backpropagation neural networks, SVMs can yield comparable accuracy using a much smaller training sample size. This is in line with the "support vector" concept that relies only on a few data points to define the classifier's hyperplane. This property has been exploited and has proved to be very useful in many of the applications we have seen thus far, mainly because acquisition of ground truth for remote sensing data is generally an expensive process."
(And more)

[14], "Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection", p.30.
"As core learners, the binary SVC and the one-class SVDD classier were used, and they were also benchmarked with neural networks in real scenarios. In general, neural networks show inferior results compared to non-linear kernel classiers, which is a direct consequence of their difculties when working with very high dimensional input samples particularly important when stacking together other information sources such as contextual or multi-temporal"

Bibliography

[1] Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

[2] A Aizerman, Emmanuel M Braverman, and LI Rozoner. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control*, 25:821–837, 1964.

[3] Hirotugu Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[4] Cesare Alippi and Manuel Roveri. An adaptive cusum-based test for signal change detection. In *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, pages 4–pp. IEEE, 2006.

[5] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L Reyes-Ortiz. Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine.

[6] Akin Avci, Stephan Bosch, Mihai Marin-Perianu, Raluca Marin-Perianu, and Paul Havinga. Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey. In *Architecture of Computing Systems (ARCS), 2010 23rd International Conference on*, pages 1–10. VDE, 2010.

[7] J. Barbič, A. Safonova, J.Y. Pan, C. Faloutsos, J.K. Hodgins, and N.S. Pollard. Segmenting motion capture data into distinct behaviors. In *Proceedings of Graphics Interface 2004*, pages 185–194. Canadian Human-Computer Communications Society, 2004.

[8] Vic Barnett and Toby Lewis. *Outliers in statistical data*, volume 3. Wiley New York, 1994.

[9] M. Basseville, I.V. Nikiforov, et al. *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs, NJ, 1993.

[10] Kristin P Bennett and Erin J Bredensteiner. Duality and geometry in svm classifiers. In *ICML*, pages 57–64, 2000.

[11] Thomas Bernecker, Franz Graf, Hans-Peter Kriegel, Christian Moennig, Dieter Dill, and Christoph Tuermer. Activity recognition on 3d accelerometer data (technical report). 2012.

[12] Robert L Brown, James Durbin, and John M Evans. Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 149–192, 1975.

[13] Fatih Camci. Change point detection in time series data using support vectors. *International Journal of Pattern Recognition and Artificial Intelligence*, 24(01):73–95, 2010.

[14] Gustavo Camps-Valls, Luis Gómez-Chova, Jordi Muñoz-Marí, José Luis Rojo-Álvarez, and Manel Martínez-Ramón. Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection. *Geoscience and Remote Sensing, IEEE Transactions on*, 46(6):1822–1835, 2008.

[15] F Chamroukhi, S Mohammed, D Trabelsi, L Oukhellou, and Y Amirat. Joint segmentation of multivariate time series with hidden process regression for human activity recognition. *Neurocomputing*, 2013.

[16] Jiun-Hung Chen. M-estimator based robust kernels for support vector machines. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 1, pages 168–171. IEEE, 2004.

[17] Tsung-Lin Cheng. An efficient algorithm for estimating a change-point. *Statistics & Probability Letters*, 79(5):559–565, 2009.

[18] Young-Sik Choi. Least squares one-class support vector machine. *Pattern Recognition Letters*, 30(13):1236–1240, 2009.

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

[20] Peter Flach. *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press, 2012.

[21] Erich Fuchs, Thiemo Gruber, Jiri Nitschke, and Bernhard Sick. Online segmentation of time series based on polynomial least-squares approximations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(12):2232–2245, 2010.

[22] Federico Girosi. An equivalence between sparse approximation and support vector machines. *Neural computation*, 10(6):1455–1480, 1998.

[23] E. Guenterberg, S. Ostadabbas, H. Ghasemzadeh, and R. Jafari. An automatic segmentation technique in body sensor networks based on signal energy. In *Proceedings of the Fourth International Conference on Body Area Networks*, page 21. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2009.

[24] Tian Guo, Zhixian Yan, and Karl Aberer. An adaptive approach for online segmentation of multi-dimensional mobile data. In *Proceedings of the Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*, pages 7–14. ACM, 2012.

[25] V. Guralnik and J. Srivastava. Event detection from time series data. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–42. ACM, 1999.

[26] Fredrik Gustafsson. The marginalized likelihood ratio test for detecting abrupt changes. *Automatic Control, IEEE Transactions on*, 41(1):66–78, 1996.

[27] Wolfgang Härdle. *Nonparametric and semiparametric models*. Springer Verlag, 2004.

[28] Zhen-Yu He and Lian-Wen Jin. Activity recognition from acceleration data using ar model representation and svm. In *Machine Learning and Cybernetics, 2008 International Conference on*, volume 4, pages 2245–2250. IEEE, 2008.

[29] J. Himberg, K. Korpiaho, H. Mannila, J. Tikanmaki, and H.T.T. Toivonen. Time series segmentation for context recognition in mobile devices. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 203–210. IEEE, 2001.

[30] Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[31] Chih-Chiang Hsu. The mosum of squares test for monitoring variance changes. *Finance Research Letters*, 4(4):254–260, 2007.

[32] Peter J Huber and Elvezio M Ronchetti. Robust statistics. 2009.

[33] Carla Inclán and George C Tiao. Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89(427):913–923, 1994.

[34] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.

[35] Y. Kawahara and M. Sugiyama. Change-point detection in time-series data by direct density-ratio estimation. In *Proceedings of 2009 SIAM International Conference on Data Mining (SDM2009)*, pages 389–400, 2009.

[36] Yoshinobu Kawahara and Masashi Sugiyama. Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining*, 5(2):114–127, 2012.

[37] E. Keogh, S. Chu, D. Hart, and M. Pazzani. An online algorithm for segmenting time series. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 289–296. IEEE, 2001.

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

[39] C. Li, SQ Zheng, and B. Prabhakaran. Segmentation and recognition of motion streams by similarity search. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 3(3):16, 2007.

[40] Song Liu, Makoto Yamada, Nigel Collier, and Masashi Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 2013.

[41] Xiaoyan Liu, Zhenjiang Lin, and Huaiqing Wang. Novel online methods for time series segmentation. *Knowledge and Data Engineering, IEEE Transactions on*, 20(12):1616–1626, 2008.

[42] Junshui Ma and Simon Perkins. Online novelty detection on temporal sequences. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–618. ACM, 2003.

[43] Junshui Ma and Simon Perkins. Time-series novelty detection using one-class support vector machines. In *Neural Networks, 2003. Proceedings of the International Joint Conference on*, volume 3, pages 1741–1745. IEEE, 2003.

[44] Markos Markou and Sameer Singh. Novelty detection: a reviewpart 1: statistical approaches. *Signal processing*, 83(12):2481–2497, 2003.

[45] Michael E Mavroforakis and Sergios Theodoridis. A geometric approach to support vector machine (svm) classification. *Neural Networks, IEEE Transactions on*, 17(3):671–682, 2006.

[46] Giorgos Mountrakis, Jungho Im, and Caesar Ogole. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3):247–259, 2011.

[47] K-R Müller, Alex J Smola, Gunnar Rätsch, Bernhard Schölkopf, Jens Kohlmorgen, and Vladimir Vapnik. Predicting time series with support vector machines. In *Artificial Neural NetworksICANN'97*, pages 999–1004. Springer, 1997.

[48] ES Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

[49] Bernhard Schölkopf, R Williamson, Alex Smola, and John Shawe-Taylor. Sv estimation of a distributions support. *Advances in neural information processing systems*, 12, 1999.

[50] Bernhard Schölkopf, Robert C Williamson, Alex J Smola, John Shawe-Taylor, and John C Platt. Support vector method for novelty detection. In *NIPS*, volume 12, pages 582–588, 1999.

[51] Erich Schubert, Arthur Zimek, and Hans-Peter Kriegel. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery*, pages 1–48, 2012.

[52] Alex J Smola, Bernhard Schölkopf, and Klaus-Robert Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11(4):637–649, 1998.

[53] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[54] Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

[55] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[56] Jun-ichi Takeuchi and Kenji Yamanishi. A unifying framework for detecting outliers and change points from time series. *Knowledge and Data Engineering, IEEE Transactions on*, 18(4):482–492, 2006.

[57] D Tax, Alexander Ypma, and R Duin. Support vector data description applied to machine vibration analysis. In *Proc. 5th Annual Conference of the Advanced School for Computing and Imaging*, pages 15–17. Citeseer, 1999.

[58] David MJ Tax. One-class classification. 2001.

[59] David MJ Tax and Robert PW Duin. Support vector domain description. *Pattern recognition letters*, 20(11):1191–1199, 1999.

[60] David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54(1):45–66, 2004.

[61] Vladimir Vapnik. Statistical learning theory. 1998, 1998.

[62] Vladimir Vapnik. *The nature of statistical learning theory*. springer, 1999.

[63] J Weston, A Gammerman, MO Stitson, V Vapnik, V Vovk, and C Watkins. Support vector density estimation. 1999.

[64] Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural computation*, 25(5):1324–1370, 2013.

[65] A.Y. Yang, S. Iyengar, S. Sastry, R. Bajcsy, P. Kuryloski, and R. Jafari. Distributed segmentation and classification of human actions using a wearable motion sensor network. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008.

[66] Jie Yin, Qiang Yang, and Jeffrey Junfeng Pan. Sensor-based abnormal human-activity detection. *Knowledge and Data Engineering, IEEE Transactions on*, 20(8):1082–1090, 2008.

[67] Shumei Zhang, Paul McCullagh, Chris Nugent, Huiru Zheng, and Matthias Baumgarten. Optimal model selection for posture recognition in home-based healthcare. *International Journal of Machine Learning and Cybernetics*, 2(1):1–14, 2011.

[68] F. Zhou, F. Torre, and J.K. Hodgins. Aligned cluster analysis for temporal segmentation of human motion. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–7. IEEE, 2008.