Chapter 13

Decision support systems for home monitoring applications

Classification of activities of daily living and epileptic seizures

Stijn Luca,[*,1]*,Lode Vuegen*[1,2]*,Hugo Van hamme*[1]*,Peter Karsmakers*[1,2]*,Bart Vanrumste*[1,2]

13.1 Introduction and overview

Home monitoring systems (HMSs) are an application of ambient intelligence that, by making use of ICT, enable home environments to become sensitive, adaptive, and responsive to the presence of people [1]. The aim of HMSs is to support the lives of people at home with respect to care and well-being and to postpone the transfer to a nursing home for people that need care. In recent years, the research to develop these services has known a rapid growth, partially due to the increasing pressure induced by the ageing population on our healthcare system.

Related to HMSs are *telemonitoring systems* which are defined as the use of telecommunication technologies to transmit data on patients health status from home to a healthcare centre [14]. Consider for example remote monitoring systems where the data of blood pressure monitors are transmitted to an external monitoring centre or emergency nurse call systems facilitating the ability to call for assistance with the push of a button. In contrast to HMSs however, telemonitoring systems do not consider the inclusion of easy-to-use technology (e.g. automated data acquisition by sensors integrated in an item of clothing) and are not adjusted to patients specific needs, nor is there any possibility for automatic adaptation when these needs are evolving.

Generally a HMS can be assigned to one of the following three different types. A first set of systems provide early diagnosis such as fall prevention methods or early diagnoses of mild-cognitive decline. A second set of systems allow patients to return

[*]corresponding author, stijn.luca@kuleuven.be - *This text is a preprint of a chapter published in the IET book Machine Learning for Healthcare Technologies (ISBN: 978-1-84919-978-0) and is subject to Institution of Engineering and Technology Copyright. A copy of the book is available at IET Digital Library*

[1] Department of Electrical Engineering, KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

[2]iMinds Future Health Department - STADIUS, KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

sooner to their homes after a hospital admittance. Consider for example systems that allow patients to do their rehabilitation exercises at home. A third and last set of systems are those that allow elderly people to postpone their transfer to a nursing home such as fall detection systems and systems that detect epileptic seizures. An essential aspect in all these systems is that real-life data is collected to build these systems. This gives more guarantees that the developed systems can be applied in practice, although this is an expensive task since (i) annotation of data leads to substantial costs; (ii) the data is often highly unbalanced due to the relevance of rare events such as falls or epileptic convulsions, requiring a lot of data to be collected; and (iii) data is often patient-specific inducing the need of training models on different patients [5].

A HMSs consist of two main components: (i) sensor technology and (ii) machine learning techniques. In this chapter the use of machine learning techniques is illustrated on data acquired by the sensors of a HMS to perform two main tasks: *activity recognition* and *novelty detection*.

The goal of activity recognition is to identify common normal activities (e.g. 'make coffee' or 'brush teeth') as they occur based on data collected by sensors. Machine learning techniques that are used to model and recognize activities include decision trees, naïve Bayes classification, Bayesian Networks, instance based learning, support vector machines (SVMs) and ensembles of classifiers that are mostly trained in a *supervised* setting where fully annotated data is needed [1].

Novelty detection aims to identify abnormal events (e.g. fall with elderly or epileptic seizures) that typically occur rarely but may indicate a crisis or an abrupt change related to health. Approaches to novelty detection include frequentist, Bayesian and information theoretic approaches, one-class support vector machines (OCSVMs) and neural networks [15]. Also the use of extreme value theory (EVT) is shown to be suitable for novelty detection [4].

The remainder of this chapter is structured as follows. In section 13.2 a tutorial on SVMs and GMMs is given. The use of these models is illustrated in a HMS where audio data is acquired to classify activities of daily living. Section 13.3 treats OCSVMs and EVT as approaches to novelty detection. The techniques are applied on an epileptic seizure detection problem. The chapter ends with some concluding remarks.

13.2 Supervised classification

In this section the classification problem is discussed in which the class $\mathscr{K}_c\,(1 \leq c \leq C)$ is estimated to which an input vector $\mathbf{x} \in \mathbb{R}^d$ belongs, e.g. the classification of handwritten digits based on pixel data. In a supervised setting this estimation is based on a training set of data containing observations whose class membership is known:

$$\mathscr{D} = \{(\mathbf{x}_i, t_i) \mid 1 \leq i \leq n\},$$

where $\mathbf{x}_i$ denote input vectors or data points in input space $\mathbb{R}^d$ and t_i denote scalar outputs or targets presenting class membership in $\{1,\ldots,C\}$.

One might divide supervised classification methods into 3 main categories: (i) *generative models*[1] that approach the classification problem by estimating a joint distribution $p(\mathbf{x},t)$ on as well inputs $\mathbf{x}$ as outputs t, (ii) *discriminative models* that only provide a model for the conditioned probabilities $p(t|\mathbf{x})$ and (iii) *discriminant functions* $f(\mathbf{x})$ that map each input $\mathbf{x}$ directly onto a class label. This section focuses on two widely known examples of models belonging to categories (i) and (iii) respectively. In particular in the following sections GMMs are used in a generative setting of classification and (2-class) SVMs are discussed as an example of a discriminant function approach where $f(\mathbf{x})$ maps each instance to one of two class labels. A typical example of a model belonging to category (ii) is given by a logistic regression model that estimates the probability of a class given an input by using a logistic function [3].

13.2.1 Gaussian mixture models for classification

In this section GMMs are introduced as a generative approach to the classification problem.

The likelihood of a GMM. The density function $p(\mathbf{x})$ of a GMM on $\mathbb{R}^d$ is given by a weighted sum of m multivariate Gaussian densities:

$$p(\mathbf{x}) = \sum_{j=1}^{m} w_j \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_i j, \boldsymbol{\Sigma}_j),$$

where $w_1, \ldots w_m$ are *mixture weights* that satisfy the constraint $\sum_{j=1}^{m} w_j = 1$ and $\mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_j \boldsymbol{\Sigma}_j)$ $(1 \leq j \leq m)$ are the density functions of d-dimensional multivariate Gaussian distributions given by:

$$\mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_j|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right),$$

with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$. Given a set of observed data points $\mathbf{x}_1, \ldots \mathbf{x}_n$ the complete set of paramaters $\boldsymbol{\lambda} = \{w_j, \mu_j, \Sigma_j | 1 \leq j \leq m\}$ can be estimated by maximizing the log likelihood function:

$$L(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \ln \left\{ \sum_{j=1}^{m} w_j \mathcal{N}(\mathbf{x}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \right\} \tag{13.1}$$

Due to the summation over j inside the logarithm in (13.1), the maximization is not analytically traceable inducing the need for a numerical algorithm as the expectation-maximization (EM) algorithm [3].

Classification with GMMs. The generative approach for classification consists of first solving the inference problem of determining the class conditional densities $p(\mathbf{x}|t)$ for each class individually. In this way a GMM is obtained for each class that is governed by a set of parameters $\boldsymbol{\lambda}_t = \{w_{tj}, \boldsymbol{\mu}_{tj}, \boldsymbol{\Sigma}_{tj} | 1 \leq j \leq m_t\}$ where the set of

[1] Generative models owe their name to the fact that they can be used to generate synthetic data points.

parameters and the number of mixture components all depend on the class described by the target variable t. The goal is then to find the maximum a posteriori (MAP) estimate $\hat{t}_{MAP}$ of the class t to which a given data point $\mathbf{x}$ belongs. Using Bayes' theorem the posterior class probabilities can be found by:

$$p(t|\mathbf{x}) = \frac{p(\mathbf{x}|t)p(t)}{p(\mathbf{x})},$$

such that:

$$\hat{t}_{MAP} := \underset{1\leq t\leq C}{\arg\max}\{p(t|\mathbf{x})\} = \underset{1\leq t\leq C}{\arg\max}\{p(\mathbf{x}|t)p(t)\} \tag{13.2}$$

One can take into account some prior belief about the class to which $\mathbf{x}$ belongs by means of the *prior distribution* $p(t)$ on the classes. Alternatively one can assume equal prior probabilities for each class reducing the estimation in (13.2) to $\hat{t}_{MAP} = \arg\max_{1\leq t\leq C}\{p(\mathbf{x}|t)\}$.

Choosing the number of components. When estimating a GMM, the number of classes has to be chosen which is not a trivial problem [3]. In a supervised setting one way to proceed is to use some of the available training data $\mathscr{D}$ to train the model with a range of values for this *hyper-parameter*. The rest of the data is split into a validation and a test set. The validation set is used to maximize performance scores (e.g. classification accuracy) while the test set is used to obtain an independent performance score to avoid over-fitting on the validation set [3]. Generally data is not abundant available inducing larger variances on the scores obtained from the validation and test data. Therefore the procedure is repeated in a K-fold cross-validation experiment where training data is partitioned into K-folds and each fold is held-out exactly once while the remaining $K-1$ folds are used for training. For a discussion on the choice of K we refer to [9]. In many application cross validations of at least 4 folds are valid choices.

13.2.2 Support Vector Machines

In this section the support vector machine (SVM) classifier is treated which is fundamentally a *two-class classifier* that assigns a data instance $\mathbf{x}$ to one of two classes presented by a target variable $t \in \{-1,1\}$. There are multiple ways to extend to multi-class SVMs. For example *one-versus-one approach* applies a 2-class SVM on all possible pairs of classes. A test instance is then assigned to that class that has the highest number of 'votes' among the classifiers [17].

The optimization problem of SVMs. The geometric problem of separation can mathematically be translated into an optimization problem minimizing the cost described by some *cost-function*. In order to find this optimal separation between the two classes a feature map $\boldsymbol{\phi} : \mathbb{R}^d \mapsto \mathbb{R}^p$ is used in an attempt to transform the geometric boundary (which is often non-linear) between the two classes in data space $\mathbb{R}^d$ to a linear boundary L in feature space (see Figure 13.1):

$$L : y(\mathbf{x}) = 0 \quad \text{with} \quad y(\mathbf{x}) = \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}) + b \quad (\mathbf{w} \in \mathbb{R}^{p\times 1}, b \in \mathbb{R}). \tag{13.3}$$

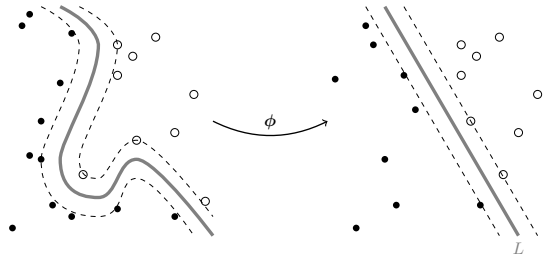


Figure 13.1: Linearisation of the decision boundary of SVMs using a feature map $\boldsymbol{\phi}$. The dashed lines indicate the hyperplanes where the margin is maximized.

The estimation of the linear boundary is performed based on a set of training examples $\mathbf{x}_i$ with corresponding target values $t_i \in \{-1,1\}$. In the ideal case this training set is linearly separable after transformation to the feature space, meaning that there exists constants $\mathbf{w} \in \mathbb{R}^{p\times 1}, b \in \mathbb{R}$ such that each training instance can be assigned to exactly one class according to the sign of $y(\mathbf{x})$ defined in (13.3). In other words one assumes that:

$$\forall 1 \leq i \leq n : t_i y(\mathbf{x}_i) > 0 \tag{13.4}$$

for some $\mathbf{w} \in \mathbb{R}^{p\times 1}, b \in \mathbb{R}$. In SVMs the *decision boundary* $L : y(\mathbf{x}) = 0$ is chosen to maximize the *margin* that is given by the smallest distance between L and any of the training instances $\mathbf{x}_i$ (Figure 13.1). In particular one is interested in constants $\mathbf{w}$ and b given by:

$$\arg\max_{\mathbf{w},b}[\min_i \left\{ \frac{|y(\mathbf{x}_i)|}{||\mathbf{w}||} \right\}] \quad \text{or} \quad \arg\max_{\mathbf{w},b}[\min_i \left\{ \frac{t_i\left(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)+b\right)}{||\mathbf{w}||} \right\}] \tag{13.5}$$

subject to the constraints (13.4). The constants $\mathbf{w}$ and $\mathbf{b}$ in (13.5) can be rescaled without changing the decision boundary $y(\mathbf{x}) = 0$ such that:

$$t_i(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)+b) = 1$$

for those instances that are closest to the decision boundary. This reduces the optimization in (13.5) to[2]:

$$\begin{aligned} &\arg\max_{\mathbf{w},b} \tfrac{1}{||\mathbf{w}||} \text{ or } \arg\min_{\mathbf{w},b} \tfrac{1}{2}||\mathbf{w}||^2 \\ &\text{subject to } t_i y(\mathbf{x}_i) = t_i(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)+b) \geq 1, \quad i = 1,\ldots,n \end{aligned} \tag{13.6}$$

Once the margin has been maximized there will be at least two instances, so-called *support vectors*, $\widetilde{\mathbf{x}}_i$ that minimize the distance to L and therefore satisfy $|y(x)| = 1$. These support vectors are lying on the *maximum margin boundaries* given by hyperplanes in feature space where the margin is geometrically maximized, see Figure 13.2(a).

[2]The factor $\frac{1}{2}$ is not necessarily but chosen for convenience when calculating derivatives of the Lagrangian in (13.11).

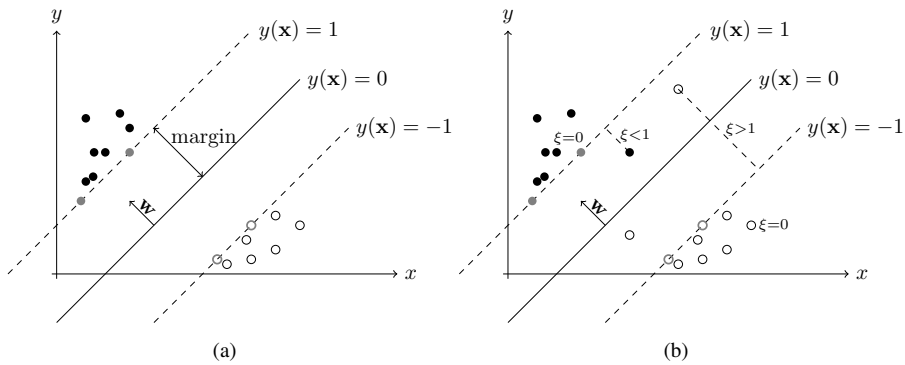


Figure 13.2: (a) Illustration of the margin of a SVM with linearly separable data. The grey points are the support vectors *lying on the* maximum margin boundaries*. (b) Illustration of the slack variables that are introduced when data is not linearly separable.*

In practice however a solution of (13.6) can not always be guaranteed as training data can be overlapping such that data points can lie at the 'wrong side' of the decision boundary. Therefore the constraints in (13.6) are weakened allowing data instances to be inside the margins using *slack variables* ξ_i. Moreover points that lie on the wrong side of the boundary are penalized in the cost function, yielding the following optimization problem which is known as the C-SVM:

$$\begin{aligned} &\arg\min_{\mathbf{w},b} \left\{ \tfrac{1}{2}||\mathbf{w}||^2 + \tfrac{C}{n}\sum_{i=1}^{n}\xi_i \right\} \\ &\text{subject to } t_i y(\mathbf{x}_i) \geq 1-\xi_i \quad \text{and} \quad \xi_i \geq 0, \quad i=1,\ldots,n. \end{aligned} \tag{13.7}$$

The slack variable ξ_i determine the error on the initial conditions $t_i y(\mathbf{x}_i) \geq 1, (1 \leq i \leq n)$ in (13.6). They are defined by $\xi_i = 0$ for support vectors or data points that are on the correct side of the margin boundaries, see Figure 13.2(b). For so-called *margin errors* lying inside the margin boundaries or at the wrong side of L one defines $\xi_i = |t_i - y(\mathbf{x}_i)|$. When $0 < \xi < 1$ they are lying inside the margin boundaries but at the correct side of L. When $\xi > 1$ the points are at the wrong side of L, see Figure 13.2(b). The parameter $C > 0$ in (13.7) determines the penalty that is put on margin errors. A lower C allows a 'softer margin', while in the limit as $C \to +\infty$ one recovers the solution for separable data as before.

From C-SVM to ν-SVM. The parameter C is rather unintuitive and there is no a priori way to select it. However a modification, called the ν-SVM is often chosen that replaces the parameter C with a parameter ν that controls the number of margin errors and support vectors as will be shown in a moment. Moreover this parametrization provides a direct link with the OCSVM that will be introduced in Section 13.3.1.

In a ν-SVM the following constrained optimization problem is solved:

$$\begin{array}{l}\arg\min\limits_{\mathbf{w},b,\rho}\left\{\frac{1}{2}||\mathbf{w}||^2-\rho\nu+\frac{1}{n}\sum\limits_{i=1}^{n}\xi_i\right\}\\ \text{subject to } \xi_i\geq 0,\rho\geq 0 \quad \text{and} \quad t_i y(\mathbf{x}_i)\geq\rho-\xi_i, \qquad i=1,\ldots,n.\end{array} \tag{13.8}$$

The maximum margin boundaries are determined by $t_i(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)+b)=\rho$ and the slack variables ξ_i determine the margin errors as before. It's not hard to realize that when ν-SVM leads to an optimum $(\mathbf{w}_0,b_0,\rho_0)$, the decision surface with coefficients $(\mathbf{w}_0,b_0)$ can equally be obtained from an optimum of the C-SVM by setting $C=\frac{1}{\rho_0}$. To see this a rescaling in the parameters $(\mathbf{w},b,\xi_i)$ in (13.8) is needed while setting $\rho=\rho_0$:

$$\overline{\mathbf{w}}=\frac{\mathbf{w}}{\rho_0},\overline{b}=\frac{b}{\rho_0},\overline{\xi_i}=\frac{\xi_i}{\rho_0} \tag{13.9}$$

such that:

$$\begin{array}{rcl}\min\limits_{\mathbf{w},b}\left\{\frac{1}{2}||\mathbf{w}||^2-\rho_0\nu+\frac{1}{n}\sum\limits_{i=1}^{n}\xi_i\right\} &=& \min\limits_{\mathbf{w},b}\left\{\frac{1}{2}||\mathbf{w}||^2+\frac{1}{n}\sum\limits_{i=1}^{n}\xi_i\right\}\\ &=& \min\limits_{\mathbf{w},b}\left\{\frac{1}{2}||\frac{\mathbf{w}}{\rho_0}||^2+\frac{1}{n\rho_0}\sum\limits_{i=1}^{n}\frac{\xi_i}{\rho_0}\right\}\\ &=& \min\limits_{\overline{\mathbf{w}},\overline{\mathbf{b}}}\left\{\frac{1}{2}||\overline{\mathbf{w}}||^2+\frac{1}{n\rho_0}\sum\limits_{i=1}^{n}\overline{\xi}_i\right\}\end{array}$$

while the constraints on $(\mathbf{w},\mathbf{b})$ in (13.8) imply the constraints (13.7) on $(\overline{\mathbf{w}},\overline{\mathbf{b}})$.

The solution of the ν-SVM optimization problem. To optimize the constraint optimization problem (13.8) the method of Lagrange multiplier is used [3]. The corresponding Lagrangian function is given by:

$$F(\mathbf{w},b,\boldsymbol{\xi},\rho)=\tfrac{1}{2}||\mathbf{w}||^2-\nu\rho+\tfrac{1}{n}\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i\left(t_i(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)+b)-\rho+\xi_i\right)-\sum_{i=1}^{n}\beta_i\xi_i-\delta\rho$$

using multipliers $\alpha_i,\beta_i\geq 0,\delta\geq 0$ subject to the conditions ('The Karush-Kuhn-Tucker' conditions):

$$\alpha_i\left(t_i(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)+b)-\rho+\xi_i\right)=0, \quad \beta_i\xi_i=0. \tag{13.10}$$

This Lagrangian F is maximized setting the first order partial derivatives to zero:

$$\begin{array}{rcl}\frac{\partial F}{\partial w_k} &=& w_k-\sum\limits_{i=1}^{n}\alpha_i t_i\boldsymbol{\phi}_k(\mathbf{x}_i)=0\Leftrightarrow w_k=\sum\limits_{i=1}^{n}\alpha_i t_i\boldsymbol{\phi}_k(\mathbf{x}_i)\\ \frac{\partial F}{\partial b} &=& \sum\limits_{i=1}^{n}\alpha_i t_i=0\\ \frac{\partial F}{\partial \xi_k} &=& \frac{1}{n}-\alpha_k-\beta_k=0\Leftrightarrow\alpha_k=\frac{1}{n}-\beta_k\\ \frac{\partial F}{\partial \rho} &=& -\nu+\sum\limits_{i=1}^{n}\alpha_i-\delta=0\Leftrightarrow\nu=\sum\limits_{i=1}^{n}\alpha_i-\delta\end{array} \tag{13.11}$$

for $1 \leq k \leq n$. Substitution in F leads to the so-called *dual representation* of the ν-SVM optimization problem:

$$\begin{aligned} F &= -\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j t_i t_j\big(\boldsymbol{\phi}(\mathbf{x}_i)\bullet\boldsymbol{\phi}(\mathbf{x}_j)\big) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \tfrac{1}{n}, \quad \sum_{i=1}^{n}\alpha_i t_i = 0, \quad \sum_{i=1}^{n}\alpha_i \geq \nu. \end{aligned} \tag{13.12}$$

In particular from (13.11), it follows that the decision function $y(\mathbf{x}) = \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}) + b$ can be written in terms of a kernel function $k(\mathbf{x},\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})\bullet\boldsymbol{\phi}(\mathbf{x}')$:

$$y(\mathbf{x}) = \sum_{i=1}^{n}\alpha_i t_i k(\mathbf{x},\mathbf{x}_i) + b.$$

Due to the conditions in (13.10) only the support vectors $\widetilde{\mathbf{x}}_i$ satisfy $\alpha_i \neq 0$ and contribute to this sum. For this reason SVMs are also called *sparse kernel machines* as the kernel function $k(\mathbf{x},\mathbf{x}')$ only has to be evaluated at a subset of the training data points reducing computation times for large data sets. Furthermore margin errors are characterised by $\xi_i > 0$ such that from (13.10) it follows that $\beta_i = 0$ and thus $\alpha_i = \frac{1}{n}$ from (13.11). As $\sum_{i=1}^{n}\alpha_i \geq \nu$ only a fraction ν of the α_i can equal $\frac{1}{n}$ such that ν is an upperbound on the fraction of margin errors as previously announced.

Kernel substitution. The dual representation (13.12) enables to work directly in terms of kernels and avoids the explicit introduction of a feature map $\boldsymbol{\phi}$, also known as the *'kernel trick'*. This allows implicitly to use feature spaces of infinite dimensionality. A commonly used kernel is given by the *Gaussian kernel*:

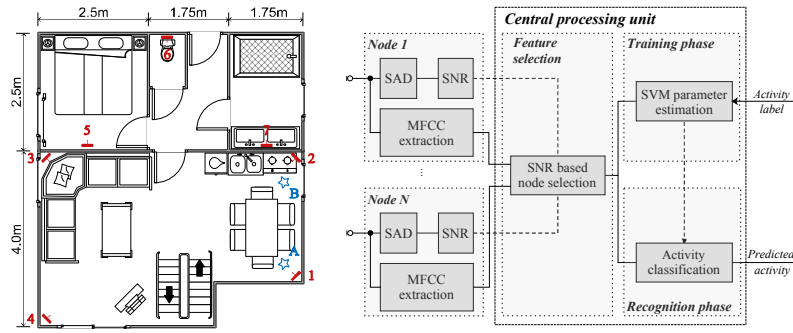
$$k(\mathbf{x},\mathbf{x}') = \exp\left(-\frac{||\mathbf{x}-\mathbf{x}'||^2}{2\sigma^2}\right) \tag{13.13}$$

which corresponds to the choice of a feature vector with infinite dimensionality and were σ denotes the so-called *kernel width*. Both σ and ν (or C) can be optimized as hyper-parameters in a cross-validation experiment similar to the procedure introduced in Section 13.2.1 for choosing the number of components in a GMM.

13.2.3 Classification of activities of daily living

In this section a supervised GMM and SVM are applied on the classification of activities of daily living from acoustic sensor data. Data is recorded in a real-life home environment equipped with seven microphone nodes. Fig. 13.3(a) shows the floor plan of the home environment together with the microphone positions. In total ten different activities of daily living were recorded during a period of three days and labelled as: 1: *'Brushing theeth'*, 2: *'Dishes'*, 3: *'Dressing'*, 4: *'Eating'*, 5: *'Preparing food'*, 6: *'Setting table'*, 7: *'Showering'*, 8: *'Sleeping'*, 9: *'Toileting'* and 10: *'Washing hands'*.

In Fig. 13.3(b) the system architecture that was used for the classification task is presented. Acoustic information is processed in blocks of 30s. Such block size



(a) (b)

Figure 13.3: (a) Floor plan of the home environment indicating the microphone positions 1 to 7. (b) The proposed system architecture for the classification of activities of daily living

corresponds to the minimal duration of activities that were observed in the data. Each block is further partitioned into frames of 25ms that overlap with 15ms. A frame is either (dominantly) generated by an "interesting" sound source or background noise sources. For each block an averaged signal-to-noise ratio (SNR) is computed as the ratio between the average energy in the interesting frames and that in the noise related frames. Hence, each 30s all nodes capture a block of data of which only that block with the highest SNR is retained and used for further processing.

Although they were initially developed for speaker and speech application Mel-Frequency Cepstral Coefficients (MFCCs) are also popular features for audio classification. They were therefore adopted in this work to form a basis on which the classifier models can work. In the setting used in this work a block contains 300 frames of 25ms. For each frame a d-dimensional MFCC feature vector $\mathbf{x}_f \in \mathbb{R}^d$ ($1 \leq f \leq 300$) is computed by retaining the d first coefficients from a cosine transformation of the log-power spectrum filtered by n_{mel} mel-filter banks [10]. In this way from each block a set of $q \leq 300$ feature vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_q\} \subset \mathbb{R}^d$ is extracted by using an energy threshold.

Both classifier models that were described in Sections 13.2.1 and 13.2.2 were validated for this task. Previous research indicated that a GMM of 10 Gaussian components with full covariance matrix is an appropriate choice for classifying activities of daily living [20]. To this end, for each frame a class dependent GMM with conditional density $p(\mathbf{x}_f|t)$ is fitted on the MFCCs feature vectors. Then, the probability that a block consisting of q frames is generated by a certain sound class is obtained as $p(\mathbf{x}_1,...,\mathbf{x}_q|t) = \prod_{f=1}^{q} p(\mathbf{x}_f|t)$. Classification of blocks could then be based on a MAP-estimation as in (13.2) assuming an uniform prior on the classes.

To apply a SVM classifier the different feature vectors of the block are described by one so-called MFCC super vector $\tilde{\mathbf{x}}_{SuVe} \in \mathbb{R}^{2d}$ defined as the first and second

Table 13.1: Mean and standard deviation computed using 4-fold cross-validation of the ADL classification accuracies for GMM, SuVe-GMM and SVM setups with different feature parameter settings. The highest obtained classification scores are marked in boldface.

n_{mel}	d	GMM			SuVe-GMM			SVM		
		8 kHz	16 kHz	32 kHz	8 kHz	16 kHz	32 kHz	8 kHz	16 kHz	32 kHz
10	7	69.6±3.3%	73.3±4.4%	73.6±5.2%	46.7±3.5%	48.3±2.6%	46.4±4.3%	68.5±5.5%	72.9±1.7%	71.4±2.8%
15	7	70.4±4.2%	73.4±4.8%	74.2±5.3%	48.0±2.2%	**52.7±4.2%**	48.2±2.0%	69.3±5.9%	72.8±4.0%	73.5±2.0%
15	14	72.8±4.8%	75.1±4.5%	**76.5±4.8%**	47.9±5.4%	50.5±5.7%	49.4±3.5%	72.8±5.1%	78.0±2.8%	76.9±2.8%
20	7	70.2±3.1%	72.8±4.9%	74.2±5.3%	47.6±8.3%	47.0±2.7%	49.5±3.5%	70.2±7.4%	72.7±0.7%	71.3±2.4%
20	14	72.7±4.4%	75.5±5.1%	73.0±4.7%	50.2±3.6%	50.0±3.1%	52.4±5.3%	69.3±2.7%	75.3±4.3%	**78.2±4.1%**

order statistics computed among the different feature vectors of a block, i.e.

$$\tilde{\mathbf{x}}_{SuVe} = \left(\frac{1}{q} \sum_{f=1}^{q} \mathbf{x}_f, \sqrt{\frac{1}{q} \sum_{f=1}^{q} (\mathbf{x}_f - \overline{\mathbf{x}}_f)^2} \right),$$

where sums and squares are component-wise defined. Also a GMM was trained using these super vectors (referred to as SuVe-GMM) in order to compare the performance of SVM and GMM when both are based on this type of feature vectors.

In Table 13.1 the mean and standard deviation of the classification accuracies (the percentage of blocks that are correctly classified) among the different type of classifiers are shown. The hyper-parameters of the GMMs and SVM are optimized in a 4-fold cross validation procedure. An one-versus-one coding scheme was used to extend the binary SVM formulation to the multi-class case.

During the experiments, the influence of the sampling frequency, the number of mel-filters n_{mel} and number of feature dimensions d on the performance are examined. As one can see, these results indicate that GMM and SVM models obtain equivalent classification accuracies and that they both outperform the SuVe-GMM setup by 20% in terms of classification accuracy. Such behaviour is typically seen when comparing generative models to discriminative functions. Given the same amount of data discriminative functions behave more robust in higher dimensional input spaces. The large difference in scores between SuVe-GMM and GMM is due to the reduction in the amount of training data while doubling the feature dimensions when using the super vector setup. In addition, these results also indicate that a sampling frequency of 16 kHz is appropriate for activity classification since lowering the sampling frequency to 8 kHz yields a decrease in accuracy while increasing to 32kHz does not improve the accuracy significantly. Therefore, SVM with a sampling frequency of 16 kHz is the preferred alternative explored in this work on this task of ADL classification.

Table 13.2 shows the confusion matrix of SVM with a sample frequency of 16kHz, 15 mel-filters and a feature dimension of 14. Most of the confusion occurs for the activities 'dishes', 'eating', 'preparing food' and 'setting table'. This seems plausible as these activities contain joint acoustic information such as scraping cutlery. In a similar way 'brussing teeth', 'dishes', 'showering', 'toileting', and

Table 13.2: SVM confusion matrix for a sample frequency of 16*kHz,* 15 *mel-filters and a feature dimension of* 14*. A classification score of* $78.0 \pm 2.8\%$ *is obtained.*

		Classified label									
		1	2	3	4	5	6	7	8	9	10
Ground truth	1	**97.9%**	2.1%	-	-	-	-	-	-	-	-
	2	1.7%	**58.6%**	6.9%	16.4%	8.6%	6.9%	-	-	-	0.9%
	3	-	0.7%	**93.5%**	3.6%	-	2.2%	-	-	-	-
	4	-	8.3%	2.9%	**77.2%**	4.9%	4.4%	1.5%	1.0%	-	-
	5	-	19.0%	3.5%	6.3%	**55.6%**	9.2%	0.7%	4.9%	0.7%	-
	6	-	6.6%	9.0%	4.1%	6.6%	**73.8%**	-	-	-	-
	7	3.1%	-	-	-	-	-	**96.9%**	-	-	-
	8	-	-	10.0%	12.5%	5.0%	-	-	**72.5%**	-	-
	9	-	-	-	-	-	-	-	-	**100%**	-
	10	4.2%	-	4.2%	-	-	-	-	-	-	**91.7%**

'washing hands' are often confused as they contain the joint acoustic signal of running water.

13.3 Novelty detection

Novelty detection is a particular example of pattern recognition that attacks the problem of identifying patterns in data that are previously unseen. It shares many similarities with anomaly detection where one also wishes to detect abnormalities, but where these may not necessarily be entirely novel, i.e. a small amount of the training data can contain outliers or anomalies. The novelty detection paradigm provides an alternative approach to strong *class imbalance* that starts from a model of normal behaviour and detects deviations from this model [15]. It is for this reason that novelty detection is also termed one-class classification where there is no explicit model for 'abnormal behaviour'. Thus in this section we start from d-dimensional training data from one class only $\mathscr{D} = \{\mathbf{x}_1, \dots \mathbf{x}_n\} \subset \mathbb{R}^d$. Statistically, the vectors $\mathbf{x} \in \mathscr{D}$ are assumed to be independent realizations of a stochastic variable X that is distributed according to a probability density function $y = p(\mathbf{x})$.

13.3.1 One-class support vector machines

A OCSVM solves an unsupervised learning problem related to a probability density estimation [17]. Instead of modelling the density of data, however, these methods aim to find a smooth boundary enclosing a region of high density. The strategy of an OCSVM is to map the training data $\{\mathbf{x}_1, \dots \mathbf{x}_n\}$ into a feature space where it can be separated from the origin with a maximal margin ρ. For this purpose the following constrained optimization problem is considered:

$$\begin{aligned} &\arg\min_{\mathbf{w},\rho} \left\{ \tfrac{1}{2}||\mathbf{w}||^2 - \rho + \tfrac{1}{n\nu} \sum_{i=1}^{n} \xi_i \right\} \\ &\text{subject to } \xi_i \geq 0 \quad \text{and} \quad y(\mathbf{x}_i) \geq \rho - \xi_i, \qquad i = 1, \dots, n. \end{aligned} \tag{13.14}$$

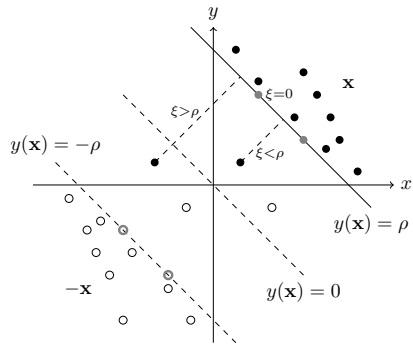


Figure 13.4: An one-class SVM pictured as a 2-class SVM on the training data and the reflected data through the origin.

where $y(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(x)$. A new instance $\mathbf{x}$ is then classified as being outside the support of the training data when $\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) - \rho \leq 0$. The optimization problem in (13.14) is very similar to the one of the ν-SVM in (13.8). In fact, rescaling the parameters in (13.14) as:

$$\mathbf{w} = \frac{\overline{\mathbf{w}}}{\nu}, \quad \rho = \frac{\overline{\rho}}{\nu}, \quad \xi_i = \frac{\overline{\xi_i}}{\nu},$$

one obtains the cost function of the ν-SVM in (13.8) where the data $\{\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}(\mathbf{x}_n)\}$ is separated from $\{-\boldsymbol{\phi}(\mathbf{x}_1), \ldots, -\boldsymbol{\phi}(\mathbf{x}_n)\}$ by the hyperplane $\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) = 0$ that passes through the origin in feature space. However OCSVMs use the maximum margin boundary $\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) = \rho$ to separate the support of the data from the rest of data space, see Figure 13.4.

Completely similar as in Section 13.2.2 the dual form can be derived by introducing the Lagrangian of the constrained optimization problem (13.14) and setting the derivatives with respect to w_i, ξ_i and ρ to zero:

$$L = -\tfrac{1}{2} \textstyle\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j t_i t_j \big(\boldsymbol{\phi}(\mathbf{x}_i) \bullet \boldsymbol{\phi}(\mathbf{x}_j)\big)$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq \tfrac{1}{\nu n}, \quad \textstyle\sum_{i=1} \alpha_i = 1.$$

The decision function in terms of the kernel function $k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x}) \bullet \boldsymbol{\phi}(\mathbf{x}')$ is now given as $y(\mathbf{x}) - \rho = \sum_{i=1}^n \alpha_i k(\mathbf{x}, \mathbf{x}_i) - \rho$. As before only the support vectors contribute to the sum. Margin errors are in this case termed outliers and the parameter ν is an upper bound on the fraction of outliers. In particular an OCSVM linearly separates the data in feature space from the origin and the choice of a Gaussian kernel (13.13) (corresponding to an infinite dimensional feature vector) ensures that this is feasible [17].

13.3.2 Extreme value theory

A main drawback of OCSVMs is the need for a choice of the parameters ν and σ. The optimal values of these parameters is depending heavily on the application such

that existing rule of thumbs generally perform suboptimal [11]. Only when examples of outliers are available the parameters can be optimized in a cross-validation experiment.

In many applications however outliers present some 'extreme' and rare behaviour. The use of EVT enables to fit a model on this class even when examples are completely absent circumventing the optimization procedure which is commonly used in SVMs. In this section we review the recent methodologies of the use of EVT for novelty detection and illustrate the methods on the detection of epileptic seizures [4, 13].

Point classification. Firstly the question is addressed whether a data point $\mathbf{x}$ is drawn from a distribution X or not. For this purpose a method is proposed that applies univariate EVT on the univariate distribution over the probability density values $p(\mathbf{x})$. The distribution Y of densities $y = p(\mathbf{x})$ is strongly related to that of X with a density function defined by:

$$q(y) = \frac{dQ}{dy}(y) \quad \text{where} \quad Q(y) = \int_{p^{-1}(]0,y[)} p(\mathbf{x}) d\mathbf{x} \tag{13.15}$$

Univariate EVT can be used to describe sets: $S_k = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ which have a typical minimal density with respect to $y = p(\mathbf{x})$. In order to avoid skewness near zero of such minimal densities, the maxima of transformed sequences $-\log(p(S_k))$ are considered:

$$m_k := \max\{-\log p(\mathbf{x}_1), \ldots - \log p(\mathbf{x}_k)\} = \max\{-\log(p(S_k)\}. \tag{13.16}$$

which corresponds to the *'extreme' vectors* with respect to X and are seen as realizations of a stochastic variable M_k. For large k, M_k follows approximately a *Gumbel distribution* with cumulative distribution function:

$$G_k(m_k) \approx \exp(-\exp(-\frac{m_k - \alpha_k}{\beta_k})) \tag{13.17}$$

where (α_k, β_k) describe respectively location and scale of the maxima related to sets S_k drawn from X.The choice of k implies a trade-off between bias and variance. A large k results in few maxima m_k that can be extracted from the training set and thus in a large estimation variance on M_k. A too small block size results in a poor estimation of the model of M_k as the approximation in (13.17) is only valid for larger k. A good compromise in our application is given by $k = 50$ [12]. In any case the validity of the approximation can visually be checked by a quantile-quantile (Q-Q) plot, graphing the empirical quantiles against the theoretical quantiles obtained from the Gumbel distribution [7].

From the training set $\mathscr{D}$ a corresponding Gumbel distribution $\hat{G}_k$ of extremes can be estimated by simulating sets S_k of length k from a kernel density estimation $y = \hat{p}(\mathbf{x})$ of $y = p(\mathbf{x})$ and obtaining the estimations $\hat{\alpha}_k$ and $\hat{\beta}_k$ of the Gumbel parameters by maximum likelihood estimation from the simulated maxima $m_k = \max\{-\log(\hat{p}(S_k)\}$ [18]. By setting a threshold on $\hat{G}_k$ a point $\mathbf{x}$ can be termed a novelty when $\hat{G}(-\log \hat{p}(\mathbf{x}))$ exceeds the threshold [3]. From a probabilistic point of

[3] A point $\mathbf{x}$ is considered as corresponding to an extreme vector of some set S_k of length k [16].

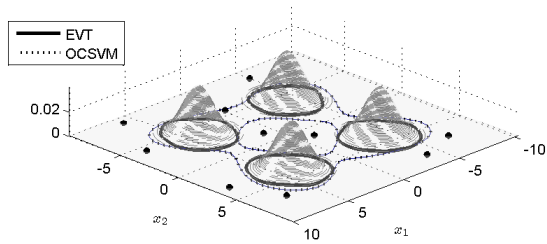


Figure 13.5: Density of a Gaussian mixture X of standard normal distributions centered at $(\pm 4, \pm 4)$. *The training instances in the abnormal class are indicated by a dot. Estimation of the support using OCSVMs and EVT is shown.*

view a threshold of 95% can be chosen corresponding to a type-I error of 5% in the classification of extremes of sets of length k.

Figure 13.5 illustrates the estimation of the support of a Gaussian mixture of standard normal distributions centered at $(\pm 4, \pm 4)$. The choice of the parameters (ν, σ) of the OCSVM is based on a cross-validation experiment using unbalanced training data consisting of 10^3 instances from the normal class and 10 instances lying in the tail of the distribution. The lack of examples from the abnormal class makes it hard for the OCSVM to estimate the correct boundary. However, EVT provides a class of models for the tail region where training data is sparse and is able to estimate the boundary better by means of extrapolation from the normal class where data is abundantly available. The support of the data then corresponds to the density contour of $\hat{p}(\mathbf{x})$ at the 95% quantile of the Gumbel distribution.

Classification of sets. We address the question of novelty detection applied on complete sets $S_k = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \subset \mathbb{R}^d$ of a specified number of k data instances that are independently drawn from some distribution. Novelty detection addresses the question whether such a set S_k of vectors is drawn from a distribution X or not. In practice S_k can e.g. present the last vector and the $k-1$ vectors observed before it such that information of the last k measurements can be combined using EVT.

In terms of statistical hypothesis testing the problem setting can be stated as:

$$\begin{array}{lcl} H_0 & : & S_k \text{ is a set of vectors drawn from the population } X \\ H_1 & : & S_k \text{ is a novel set with respect to } X \end{array}$$

From the point of view of hypothesis testing, it is clear that for $k > 1$ the problem is related to one of multiple testing. Indeed, for $k > 1$ the probability to make at least one false positive when testing each $\mathbf{x}_i \in S$ is given by:

$$P(\text{false positive}) = 1 - (1-\alpha)^k > \alpha,$$

where α denotes the probability on a false positive when testing a single $\mathbf{x}_i$. As k gets larger the probability of a false alarm drastically increases. When e.g. $k = 5$ and $\alpha = 5\%$, then P(false positive)=26%. The use of EVT enables to obtain the correct boundary of normality corresponding with the significance level α.

In order to classify such sets it is desired to fuse different types of information of S_k in order to build a classification model. The use of Poisson point processes (PPPs) allows us to do this in a very natural way as these models will allow us to fuse three different types of information of S_k given some threshold u: (i) the maximal exceedance m_k of $-\log p(S_k)$ above u (ii) the mean exceedance v_k of $-\log p(S_k)$ above u, and (iii) the number of exceedances n_k of $-\log p(S_k)$ above u. The distributions of the corresponding random variables M_k, V_k and N_k can be obtained by applying the PPP approach.

This approach of EVT states that the number of exceedances in $-\log p(S_k)$ above some high threshold u can be approximated by a *Poisson distribution* for large k, with a rate λ_k that can be parametrised in terms of the Gumbel parameters (α_k, β_k):

$$\lambda_k = \exp\left(\frac{u-\alpha_k}{\beta_k}\right) \tag{13.18}$$

The choice of u implies the same trade-off as the choice of k, a too large u results in a large estimation variance on the parameters $(\lambda_k, \alpha_k, \beta_k)$ while a too low u implies a poor approximation by the Poisson distribution. Compromises are described by rule of thumbs such as *Van Kerm's rule* stating that $u \approx \min\{\max\{2.5\overline{x}, q_{98}\}, q_{97}\}$ where $\overline{x}, q_{98}, q_{97}$ denote empirical estimates of mean and quantiles at $0.98, 0.97$ respectively using a sample drawn from $-\log p(X)$ [2]. As before, a kernel density estimation $y = \hat{p}(\mathbf{x})$ of $y = p(\mathbf{x})$ can be obtained from the training set $\mathscr{D}$ from which a number of n_b sets S can be simulated. When one observes m exceedances $z_i - u$, $z_i = -\log \hat{p}(\mathbf{x}_i)$ among these sets, the EVT parameters λ_k, α_k and β_k can be estimated by maximizing the Poisson process log-likelihood [7]:

$$-n_b \exp\left(\frac{u-\alpha_k}{\beta_k}\right) - m\log\beta_k - \sum_{i=1}^{m}\left(\frac{z_i-u}{\beta_k}\right) \tag{13.19}$$

Now, according to EVT, M_k (equation (13.16)) follows a Gumbel distribution with location α_k and scale β_k, N_k a Poisson distribution with rate λ_k and the exceedances $-\log(p(S_k)) - u$ an exponential distribution with scale β_k. The latter implies that given a number of exceedances n_k the variable V_k follows an *Erlang distribution* with shape-parameter n_k and rate parameter $\frac{n_k}{\beta_k}$. With respect to each of the distributions M_k, N_k and V_k, a set S_k can be evaluated by means of a cumulative probability score that we respectively denotes as $\chi_g(S_k)$, $\chi_p(S_k)$ and $\chi_e(S_k)$ (the subindices refer to the underlying distributions: Gumbel, Poisson and Erlang). These scores can be combined into one *novelty score* of S_k using a generalized mean:

$$\overline{\chi}_r(S_k) = \left(\frac{1}{3}(\chi_p(S_k)^r + \chi_e(S_k)^r + \chi_g(S_k)^r)\right)^{1/r} \tag{13.20}$$

Depending on the application one can choose an appropriate r. When $r \mapsto 0$ one obtains a geometric mean while for $r \mapsto -\infty$ and $r \mapsto +\infty$ one gets the minimal and maximal score respectively. Furthermore $\overline{\chi}_r(S_k)$ is increasing as a function of r such that depending on the choice of r the sensitivity of the algorithm is influenced. A choice of $r = +\infty$ leads to a novelty system that gives an alarm when at least one cumulative probability exceeds a threshold and therefore implies maximal sensitivity

but possible higher false alarm rates. For $r = -\infty$ all cumulative probabilities have to exceed a threshold implying less false alarms and thus generally lower sensitivity. All other choices are situated between these two extremes.

13.3.3 Epileptic seizure detection

In this section a case study in healthcare is considered using a data set of acceleration data collected from movements of patients suffering from epilepsy [6]. The acceleration data was recorded during several nights using four 3D acceleration sensors that are attached to the extremities of 7 patients with hypermotor seizures, all between the age of 5 and 16 years. Hypermotor seizures are epileptic convulsions that are marked by a strong and uncontrolled movement of the arms and legs that can last from a couple of seconds to some minutes. Due to the heavy movement, the patient can injure himself during the seizure, which increases the need for an alarm system, with a high detection rate.

Movement events E_s are extracted from the data set using an energy threshold. Denote the acceleration vectors in these events as $E_s = \{\mathbf{a}_{tl} | 1 \leq t \leq T, 1 \leq l \leq 4\}$ where the indices refer to the time index and the limb respectively (1=left arm, 2=right arm, 3=left leg, 4=right leg). A feature analysis [6] identifies 3 important features: (i) the movement length $f_1 = |E_s| = T$, (ii) the average energy in a movement:

$$f_2 = \frac{1}{T} \sum_{t,l} \|\mathbf{a}_{tl}\|^2,$$

and (iii) the average of the maximal energy in an arm movement:

$$f_3 = \frac{1}{T} \sum_{t} \max\{\|\mathbf{a}_{t1}\|^2, \|\mathbf{a}_{t2}\|^2\}.$$

The features are calculated on 50% overlapping sliding windows containing 125 samples [12] which are randomly subsampled to obtain sets S_k of fixed length $k = 50$ containing 3-dimensional data instances $\mathbf{x}_i = (f_1^i, f_2^i, f_3^i)$, $1 \leq i \leq 50$ on which the EVT algorithm for the classification of sets can be applied. The validity of the Gumbel model for $k = 50$ can be assessed by means of quantile-quantile (Q-Q) plots [12].

In an EVT approach a kernel density estimation is performed to estimate the distribution X representing non-seizure movements and the related EVT parameters α_k, β_k and λ_k for $k = 50$. The kernel width is set to $H = n^{-2/7}\hat{\Sigma} \in \mathbb{R}^{3\times 3}$ according to Scott's rule of thumb [18] where n denotes the number of data points in the training set and $\hat{\Sigma}$ the sample covariance matrix. Sets are classified by using the novelty score (13.20) while setting $r = -\infty$ and thresholding at 95%. This allows to minimize the false alarm rate in a 10-fold cross validation experiment while the detection rate stayed at a high level. To evaluate our method the sensitivity (SS) and positive predictive value (PPV) is used:

$$SS = \frac{TP}{FP+FN}, \quad PPV = \frac{TP}{TP+FN}$$

Table 13.3: Means and standard deviations of SS and PPV in a 10*-fold cross-validation experiment for patients 1-7 based on an OCSVM and an EVT classifier.*

	OCSVM							EVT					
Pat.	SS			PPV			σ	SS			PPV		
1	100.0	±	0.0	31.66	±	16.08	0.01	100.0	±	0.0	52.8	±	35.9
2	100.0	±	0.0	37.90	±	10.22	0.01	100.0	±	0.0	71.8	±	18.9
3	100.0	±	0.0	40.19	±	11.17	0.14	100.0	±	0.0	64.7	±	21.5
4	100.0	±	0.0	17.62	±	5.33	0.56	70.0	±	25.8	40.5	±	32.2
5	64.44	±	10.21	19.12	±	36.94	0.81	13.3	±	11.5	15.8	±	13.1
6	100.0	±	0.0	39.04	±	24.40	0.01	100.0	±	0.0	69.6	±	24.6
7	100.0	±	0.0	40.07	±	17.03	0.09	100.0	±	0.0	52.6	±	12.4

where the number of seizures that is detected is denoted as TP ('true positive') and the number that are not detected as FN ('false negatives') while FP ('false positives) denotes the number of normal movements that triggered an alarm, see table 13.3.

The use of PPPs for epileptic seizure detection seems appropriate as it is indeed plausible that a typical epileptic convulsion does not result in one very high excess in the acceleration data but to multiple exceedances with a high mean excess. Only for patient 5 a low PPV score was obtained due to the fact that for this patient seizures seemed less 'extreme' and thus less excesses were observed [6]. To illustrate this fact, consider the two movements of patient 2 shown in Fig. 13.6. As well the normal movement as the seizure contain extremes that exceed the threshold t determined by

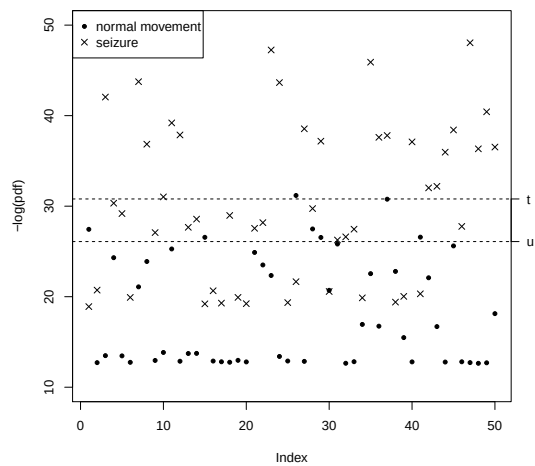


Figure 13.6: Plot of the log-densities $-\log(p(\mathbf{x}_i)), 1 \leq i \leq 50$ *of a normal movement and a seizure. The threshold* t *corresponds to the* 95% *quantile of the Gumbel distribution on* M_k *and* u *denotes the threshold as in* (13.18) *estimated by Van Kerm's rule of thumb.*

the 95% quantile of the Gumbel distribution of M_k. However the movements in the seizure are clearly more violent than the normal movement. Because the number of exceedances above u is high for each movement the scores $\chi_p(S_k)$ exceed 99% for both movements. However there is a clear difference between the scores $\chi_e(S_k)$ that describe the mean excesses that are given by 80.47% and 99.99% for the normal movement and seizure respectively.

As discussed in Section 13.3.1 an alternative approach to this novelty detection problem is an OCSVM classifier. To this end, features are extracted from complete movements such that each movement is represented by 1 feature vector. To make a consistent comparison with the EVT-method the same features and randomizations during the 10-fold cross validation are chosen. The parameter ν was set to 0.05 in accordance with the 95% threshold on the novelty scores based on the EVT-method and performance scores were optimized with respect to the kernel width σ varying over the range $]0,10]$ with a step size of 0.01. Results are shown in table 13.3. The PPV scores of patients $1-4$ and $6-7$ are maximized while the SS scores are kept at 100%. The EVT-method is able to outperform the SVM approach in 5 of the 7 patients with a mean increase in PPV of 24.5%. For patient 5 it is possible to obtain a higher SS score and PPV score in comparison with our EVT-method by setting $\sigma = 0.81$. For this patient the SVM method was able to outperform the EVT method, although in contrast to the EVT approach the hyper-parameters of SVM were tuned using data from the seizures.

13.4 Conclusion

The focus in this chapter was on activity recognition and novelty detection that are at the core of HMS technologies.

Short tutorials were provided on GMMs and SVMs for supervised classification tasks. When applying these methods on a real-life application of classifying activities of daily living, it was found that the discriminative approach of SVM outperformed the GMM. The use of these supervised methods require expert interaction for labelling and therefore result in a substantial cost in practice. This implies the need for semi-supervised methods, where as well labelled as unlabelled data is used. Existing attempts are not adapted for their use in HMS environments where scalability (being able to roll-out a system with a high number of users) and re-usability (being able to apply the same model on different persons) are ongoing challenges [8, 19].

For novelty detection OCSVMs and EVT are applied on the detection of epileptic seizures using accelerometer data. OCSVMs have the disadvantage to depend on several hyper-parameters that need to be tuned in a cross-validation experiment requiring data from the abnormal class. However EVT is a field in statistics that is especially developed to form models of data that are situated away from the modes of a distribution and which can be adapted to circumvent the tuning of several parameters. The scarcity of the occurrence of abnormalities in many applications of HMSs requires an unusual high accuracy of novelty detection algorithms to overcome a

high false alarm rate. Therefore combining several types of information using rich models (as e.g. PPPs) is a must in order to limit the number of false alarms.

Bibliography

[1] ACAMPORA, G., COOK, D., RASHIDI, P., AND VASILAKOS, A. A survey on ambient intelligence in health care. *Proceeding of the IEEE 101*, 12 (2013), 2470–2494.

[2] ALFONS, A., AND TEMPL, M. Estimation of social exclusion indicators from complex surveys: The R package laeken. *Journal of Statistical Software 54*, 15 (2013), 1–25.

[3] BISHOP, C. *Pattern Recognition and machine learning*. Springer, New York, USA, 2006.

[4] CLIFTON, D., HUGUENY, S., AND TARASSENKO, L. Novelty detection with multivariate extreme value statistics. *Journal of Signal Processing Systems 65* (2011), 371–389.

[5] CROONENBORGHS, T., LUCA, S., KARSMAKERS, P., AND VANRUMSTE, B. Healthcare decision support systems at home. In *Artificial Intelligence Applied to Assistive Technologies and Smart Environments: Papers from the AAAI-14 Workshop* (2014), B. Bouchard, A. Bouzouane, S. Giroux, A. Mihailidis, and S. Guillet, Eds., pp. 9–10.

[6] CUPPENS, K., KARSMAKERS, P., VAN DE VEL, A., BONROY, B., MILOSEVIC, M., LUCA, S., CEULEMANS, B., LAGAE, L., VAN HUFFEL, S., AND VANRUMSTE, B. Accelerometer based home monitoring for detection of nocturnal hypermotor seizures based on novelty detection. *IEEE Journal of Biomedical and Health Informatics 60*, 2 (2013), 89–96.

[7] EMBRECHTS, P., KLÜPPELBERG, C., AND MIKOSCH, T. *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin, 1997.

[8] GUAN, D., YUAN, W., LEE, Y.-K., AND GAVRILOV, L. Activity recognition based on semisupervised learning. In *13th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications* (2007), IEEE, pp. 469–475.

[9] HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. *The Elements of Statistical Learning*. Springer, New York, 2001.

[10] HUANG, X., ACERO, A., AND HON, H.-W. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st ed. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001.

[11] JAAKKOLA, T., DIEKHANS, M., AND HAUSSLER, D. Using the fisher kernel method to detect remote protein homologies. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology* (1999), AAAI Press, pp. 149–158.

[12] LUCA, S., KARSMAKERS, P., CUPPENS, K., CROONENBORGHS, T., VAN DE VEL, A., CEULEMANS, B., LAGAE, L., VAN HUFFEL, S., AND VANRUMSTE, B. Detecting rare events using extreme value statistics applied to epileptic convulsions in children. *Journal of Artificial Intelligence In Medicine 60*, 2 (2014), 89–96.

[13] LUCA, S., KARSMAKERS, P., AND VANRUMSTE, B. Anomaly detection using the poisson process limit for extremes. In *IEEE International Conference on Data Mining* (2014), R. Kumar, H. Toivonen, J. Pei, Z. H., and X. Wu, Eds., pp. 370–379.

[14] PARÉ, G., JAANA, M., AND SICOTTE, C. Systematic review of home telemonitoring for chronic diseases: The evidence base. *Journal of the American Medical Informatics Association 14*, 3 (2007), 269–277.

[15] PIMENTEL, M. A. F., CLIFTON, D., CLIFTON, L., AND TARASSENKO, L. A review of novelty detection. *Signal Processing 99* (2014), 215 – 249.

[16] ROBERTS, S. Novelty detection using extreme value statistics. *IEE Proceedings on Vision, Image and Signal processing 146*, 3 (1999), 124–129.

[17] SCHÖLKOPF, B., AND SMOLA, A. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, London, 2002.

[18] SCOTT, D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley and Sons, New York, 1992.

[19] STIKIC, M., LARLUS, D., EBERT, S., AND SCHIELE, B. Weakly supervised recognition of daily life activities with wearable sensors. *IEEE Transactions on Patterns Analysis and Machine Intelligence 33*, 12 (2011), 2521–2537.

[20] VUEGEN, L., VAN, B., BROECK, D., KARSMAKERS, P., HAMME, H. V., AND VANRUMSTE, B. Automatic monitoring of activities of daily living based on real-life acoustic sensor data: a preliminary study. In *Workshop on speech and language processing for assistive technologies* (2013), Association for Computational Linguistics (ACL), pp. 113–118.