Early and reliable event detection using proximity space representation

Maxime Sangnier, Jérôme Gauthier, A. Rakotomamonjy

▶ **To cite this version:**

Maxime Sangnier, Jérôme Gauthier, A. Rakotomamonjy. Early and reliable event detection using proximity space representation. ICML 2016 - 33rd International Conference on Machine Learning, Jun 2016, New York, United States. pp.2310-2319. cea-01843181

HAL Id: cea-01843181
https://cea.hal.science/cea-01843181

Submitted on 9 May 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Early and Reliable Event Detection Using Proximity Space Representation

Maxime Sangnier MAXIME.SANGNIER@TELECOM-PARISTECH.FR
LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France

Jérôme Gauthier JEROME.GAUTHIER@CEA.FR
LADIS, CEA, LIST, 91191, Gif-sur-Yvette, France

Alain Rakotomamonjy ALAIN.RAKOTO@INSA-ROUEN.FR
Normandie Université, UR, LITIS EA 4108, Avenue de l'université, 76801, Saint-Etienne-du-Rouvray, France

Abstract

Let us consider a specific action or situation (called event) that takes place within a time series. The objective in early detection is to build a decision function that is able to go off as soon as possible from the onset of an occurrence of this event. This implies making a decision with an incomplete information. This paper proposes a novel framework that i) guarantees that a detection made with a partial observation will also occur at full observation of the time-series; ii) incorporates in a consistent manner the lack of knowledge about the minimal amount of information needed to make a decision. The proposed detector is based on mapping the temporal sequences to a landmarking space thanks to appropriately designed similarity functions. As a by-product, the framework benefits from a scalable training algorithm and a theoretical guarantee concerning its generalization ability. We also discuss an important improvement of our framework in which decision function can still be made reliable while being more expressive. Our experimental studies provide compelling results on toy data, presenting the trade-off that occurs when aiming at accuracy, earliness and reliability. Results on real physiological and video datasets show that our proposed approach is as accurate and early as state-of-the-art algorithm, while ensuring reliability and being far more efficient to learn.

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

1. Introduction

Early detection of temporal events is a valuable ability for automatic systems that serve in many widespread fields. Common instances of such applications are security (*e.g* video surveillance, attack detection and earthquake warning), healthcare (*e.g* heart failure detection and protection of the elderly) and entertainment (*e.g* recognition of gestures and gaming). Concretely, early detection is the capability of detecting as soon as possible an occurrence of the event of interest (let us say a heart disorder) during an online sequential analysis of a time-series (an electrocardiogram in this example). This implies making a decision with the incomplete observation of an occurrence, that is a partial information. Admitting that the temporal event of interest is of finite duration, our objective is to build a detector that is able to make a correct decision as soon as an occurrence appears and obviously before it ends.

Early decision systems got a growing interest from the machine learning community in the last years. This is true for early classification (Xing et al., 2009; 2012; Parrish et al., 2013) and detection (Hoai & De la Torre, 2012; 2014). In these works, two points are noteworthy.

The first important point is *earliness*. The detector is learned so as to make a decision with partial observations but without knowing exactly the sufficient amount of information to collect. A tempting way to achieve earliness is to force partial observations to be well recognized (Ellis et al., 2013; Hoai & De la Torre, 2014). This is quite computationally demanding and more importantly a simplistic way to handle the lack of knowledge about the minimal amount of information required to make a decision. Indeed, such a procedure implies considering partial observations as occurrences of the event. However, some of those incomplete observations should not be considered as positive events since not enough information has been collected yet. This may confuse the learning of the recognition system.

The second important concept is *reliability*. It is defined in a probabilistic way in Parrish et al. (2013): *the probability that a prediction given incomplete information is the same as the one given the complete information*. This property is essential for early systems since it guarantees the consistency between the decisions with a partial and a full observation. Such a consistency is a mandatory property in medical applications or other security applications.

In this work, we describe a novel and general framework, nicknamed SimpleED, to build a *reliable* and *early* nonlinear detector of temporal events. Here, we consider reliability in a deterministic way, that is, we ensure that the decision with a partial observation is identical to the one achieved with the full sequence. However, achieving these two properties come inherently with the price of a small trade-off in detection performances.

For this purpose, we assume that the events are characterized by discriminative frames, where a frame corresponds to a single unit in the temporal sequence (an image for a video sequence or a frame for an audio recording). Such a frame-based approach is well suited to a sequential analysis, since at each time step a new frame is collected.

Moreover, this framework is naturally connected to multiple-instance learning (MIL). The MIL paradigm is aimed at labeling a bag of instances based on these ones, but without knowing beforehand which instance is discriminative. This ambiguity is inherent to MIL and makes it a relevant problem (Keeler et al., 1991; Dietterich et al., 1997; Wang & Zucker, 2000). Our frame-based approach makes a link between the MIL ambiguity and the lack of knowledge concerning the sufficient amount of information to collect to make an early decision. Thus, and similarly to the MIL paradigm, this minimal amount is obtained (and then the event occurrence can be detected) as soon as a discriminative frame appears. As such, we have build our framework from the MIL work by Chen et al. (2006) and we have significantly extended it by exploiting its full potential for sequential decision and by integrating earliness and reliability properties.

In summary, we make the following contributions in the context of early detection of temporal events (reflecting the outline of the paper):

- we propose a novel framework for learning an early detector which achieves a guaranteed and deterministic reliability;
- this framework handles in a consistent manner the lack of knowledge inherent to early detection by taking inspiration from the MIL literature. It is based on appropriately designed similarity measures which nicely embed the sequences of frames. Owing to that, we succeed in building an early detector framework without the need of enumerating all partial observations;
- even with the special features of early detection, our framework comes with Rademacher-based generalization guarantees, obtained from the work by Kakade et al. (2009). To the best of our knowledge, this is the first theoretical analysis concerning generalization of an early detector and it comes as an important by-product of our learning framework;
- while the model ensuring reliability is achieved by restricting the expressivity power of the classifier, we show that a large class of reliable models can actually be built by relaxing these constraints and by a slight modification of the decision function. Experimental results show that these relaxed models are competitive both in terms of earliness and accuracy while still ensuring reliability.

Besides these novelties, the last section of this paper is devoted to a detailed comparison between our framework and the works by Chen et al. (2006) and Hoai & De la Torre (2014). The numerical experiments include physiological and video datasets.

2. Related work

The general topic of event detection has been widely explored in several fields like computer vision (Gorelick et al., 2007) and disease outbreak (Neill et al., 2005). Yet, early detection has just recently appeared in the machine learning community (Hoai & De la Torre, 2012; 2014), resulting in the method called maximum margin early event detector (MMED). In this work, the authors extend structured output SVM (Tsochantaridis et al., 2005) to handle the sequential nature of time-series and early detection by augmenting the training sequences with all the partial observations. Then reliability is touched upon thanks to explicit constraints that promote an increase of the decision during the analysis of an event occurrence. This approximate growth of the decision function conveys the following idea: the more information is collected, the *more positive* the decision should be. However, the learned detector is not deterministically reliable. In addition, by nature, this approach is rather computationally demanding due to training example augmentation and the structured output learning framework.

Early event detection has a strong relation with early classification of time-series, which is an active field of research since the early attempts by Rodriguez & Alonso (2002) and latter by Xing et al. (2009; 2012). Rodriguez & Alonso (2002) built an early classifier by boosting weak learners based on simple predicates like *the time-series increases in this region* or *the time-series stays in this region*. On the contrary, Xing et al. (2012) have developed an early classifier based on a nearest-neighbor technique. More recently, a framework to classify temporal sequences as soon as pos-

sible and with a predefined probabilistic reliability has been introduced in (Parrish et al., 2013). By estimating the first moments of the conditional density of the complete data given the incomplete data, Parrish et al. are able to classify an event with low latency and high reliability. This approach is based on some hypotheses on the probability distribution of the data, that enable them to derive an early decision function.

The MIL paradigm is naturally suited to recognition of temporal sequences, since it deals with discovering the discriminative instances in a bag (and so the discriminative frames in a sequence). Thus, MIL has been recently applied to early recognition of human actions in a probabilistic setting (Ellis et al., 2013). Like (Hoai & De la Torre, 2014), the system proposed by Ellis et al. tackles earliness through augmented training sequences (partial observations) and does not focus on reliability.

Compared to all above-mentioned works, as well as a preliminary study of ours (Sangnier et al., 2015), our approach is based on a framework which makes no hypotheses on the data distribution while providing guaranteed reliability of the decision. In addition, the learning problem is simple and does not need partial training sequences, making it more scalable than competitors. Finally, it comes with a generalization bound on the expected loss of the detector.

3. Theoretical framework for early detection

3.1. Problem definition

In this work, the time is discretized and embodied by the subscript $t \in \{1, 2, \dots, T\}$ with T being a predefined upper-bound. At each t, a frame is represented by a feature vector $\boldsymbol{x_t}$, coming from some set $\mathcal{S}$. A temporal sequence is a time-feature representation $\mathbf{X}_{1..T} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_t}, \dots, \boldsymbol{x_T})$ and we define $\mathcal{X}_T$ the set of these sequences. Shorter sequences such as $\mathbf{X}_{1..t}$, with $t < T$, are considered elements of $\mathcal{X}_T$ through zero-padding.

The aim of this study is to build a real-valued decision function $f\colon \mathcal{X}_T \to \mathbb{R}$ such that $f(\mathbf{X}_{1..t})$ predicts the nature of the full sequence $\mathbf{X}_{1..T}$, given the partial observation $\mathbf{X}_{1..t}$. For a detection threshold $b \in \mathbb{R}$, $f(\mathbf{X}_{1..t}) \geq b$ claims that the sequence $\mathbf{X}_{1..T}$ is an occurrence of the event of interest (this is a detection), while $f(\mathbf{X}_{1..t}) < b$ means that the sequence $\mathbf{X}_{1..T}$ does not represent the event or that the detector did not collect enough information to make a detection. This is the default state.

The problem of early detection is to detect an event occurrence as soon as possible. Concretely, we shall produce a decision $f(\mathbf{X}_{1..t}) \geq b$ with the shortest partial observation $\mathbf{X}_{1..t}$ (the smallest t), only when $\mathbf{X}_{1..T}$ represents the event. In practice, such a detector is used in a sequential way. A decision is computed at each time step: $f(\mathbf{X}_{1..1})$, $f(\mathbf{X}_{1..2})$, …, $f(\mathbf{X}_{1..t})$. When $f(\mathbf{X}_{1..t}) \geq b$, the analysis is interrupted and a notification of detection is thrown. Note that this sequential analysis can be compacted by defining $g(\mathbf{X}_{1..T}) \triangleq \max_{1\leq t\leq T} f(\mathbf{X}_{1..t})$. This means that $\mathbf{X}_{1..T}$ is declared as an event occurrence if and only if $g(\mathbf{X}_{1..T}) \geq b$.

The main aim of this paper is to propose a framework where f is *early* and *reliable*, that is $f(\mathbf{X}_{1..T})$ and $g(\mathbf{X}_{1..T})$ produce the same decisions. For this purpose, let us define formally the notion of reliability. The definition we provide is inspired by the probabilistic one by (Parrish et al., 2013) but differs in that it is deterministic.

Definition 3.1 (Reliability) *A decision function f is said reliable with respect to a detection threshold $b \in \mathbb{R}$ if:*

$$\forall \mathbf{X}_{1..T} \in \mathcal{X}_T, f(\mathbf{X}_{1..T}) < b\colon \max_{1\leq t\leq T} f(\mathbf{X}_{1..t}) < b.$$

With this definition, a reliable couple (f, b) satisfies $\operatorname{sign}\left(\max_{1\leq t\leq T} f(\mathbf{X}_{1..t}) - b\right) = \operatorname{sign}\left(f(\mathbf{X}_{1..T}) - b\right)$. In other words, this definition tells that if the detector during the sequential prediction phase outputs a positive label, then the observation of the full sequence $\mathbf{X}_{1..T}$ would also lead to a positive detection. In addition, if the detector does not early trigger then it wont trigger when observing $\mathbf{X}_{1..T}$.

3.2. Inducing reliable models

Consider a set of training sequences $\left\{\left(\mathbf{X}^{(i)}_{1..T}, y_i\right)\right\}_{1\leq i\leq n}$, where the label y_i is equal to $+1$ when the sequence $\mathbf{X}^{(i)}_{1..T}$ is an event occurrence, or to -1 otherwise. Our objective is to learn from these examples a decision model $f(\cdot)$ that is early and reliable. In what follows, we expose how we induce model reliability.

The proposed framework is based on a similarity measure $k\colon \mathcal{X}_T \times \mathcal{S} \to \mathbb{R}$ which, in a nutshell, measures the similarity of a sequence $\mathbf{X}_{1..t}$ to a single frame representation $\boldsymbol{p}$. Typically, this frame representation $\boldsymbol{p}$, called from now on a landmark, is expected to be related to a discriminative frame. Note that the similarity measure k performs some sort of pooling over time, since sequences of different lengths can be compared to a single frame $\boldsymbol{p}$. Owing to this measure k and a set of landmarks $\{\boldsymbol{p}_j\}_{j=1}^m$, the decision function f is defined as: $f(\mathbf{X}_{1..t}) = \langle \mathbf{w} \mid \psi(\mathbf{X}_{1..t})\rangle_{\ell_2}$, where $\mathbf{w} \in \mathbb{R}^m$ and $\psi\colon \mathcal{X} \to \mathbb{R}^m$ is a map such that $\psi(\mathbf{X}_{1..t}) = (k(\mathbf{X}_{1..t}, \boldsymbol{p_1}), \dots, k(\mathbf{X}_{1..t}, \boldsymbol{p_m}))$. The reason for this choice stands on some theoretical arguments that are exposed in the sequel, but also on very intuitive ones. Indeed, as the landmarks are supposed to be discriminative for the task at hand, a sequence $\mathbf{X}_{1..t}$ that exhibits strong similarity with a landmark is expected to be detected reliably and as soon as this resemblance is made clear.

Let us now introduce formally the few assumptions required to build a reliable model $f(\cdot) = \langle \mathbf{w} \mid \psi(\cdot) \rangle_{\ell_2}$.

Proposition 3.1 *Let $k\colon \mathcal{X}_T \times \mathcal{S} \to \mathbb{R}$ be a similarity measure and $\mathbf{w} \in \mathbb{R}^m$. If $\mathbf{w} \succcurlyeq 0$ and if k is a non-decreasing time-dependent function: $\forall (\boldsymbol{p}, \mathbf{X}_{1..T}) \in \mathcal{S} \times \mathcal{X}_T$,*

$$\Big[t_1 \leq t_2\Big] \Rightarrow \Big[k(\mathbf{X}_{1..t_1}, \boldsymbol{p}) \leq k(\mathbf{X}_{1..t_2}, \boldsymbol{p})\Big],$$

then $f(\cdot) = \langle \mathbf{w} \mid \psi(\cdot) \rangle_{\ell_2}$ is reliable with respect to any detection threshold.

Proof With these assumptions, $f(\mathbf{X}_{1..t_1}) \leq f(\mathbf{X}_{1..t_2})$ when $t_1 \leq t_2$. Thus, $\max_{1\leq t\leq T} f(\mathbf{X}_{1..t}) = f(\mathbf{X}_{1..T})$ and f is reliable. ■

This proposition tells us that by imposing non-negative weights ($\mathbf{w} \succcurlyeq 0$) and by appropriately choosing the similarity measure k, the resulting detector f is reliable.

In this work, we propose to build a non-decreasing similarity measure k according to the following recipe. Consider a frame-to-frame proximity function $q\colon \mathcal{S} \times \mathcal{S} \to \mathbb{R}$, for instance $q(\boldsymbol{x_{t'}}, \boldsymbol{p}) = \langle \boldsymbol{x_{t'}}, \boldsymbol{p} \rangle_{\mathcal{S}}$, or $q(\boldsymbol{x_{t'}}, \boldsymbol{p}) = \exp\left(-\gamma \|\boldsymbol{x_{t'}} - \boldsymbol{p}\|_{\mathcal{S}}^2\right)$, where $\gamma \geq 0$, depending whether the set $\mathcal{S}$ admits an inner product or a norm. Then, by pooling the past proximity values, for example thanks to an ℓ_r-norm: $k(\mathbf{X}_{1..t}, \boldsymbol{p}) = \left(\sum_{t'=1}^{t} |q(\boldsymbol{x_{t'}}, \boldsymbol{p})|^r\right)^{\frac{1}{r}}$, the resulting similarity function k is non-decreasing and can be used to learn a reliable detector. When $r \to +\infty$ and $q(\cdot, \cdot)$ is Gaussian, this procedure returns the function:

$$k(\mathbf{X}_{1..t}, \boldsymbol{p}) = \exp\left(-\gamma \min_{1\leq t'\leq t} \|\boldsymbol{x_{t'}} - \boldsymbol{p}\|_{\mathcal{S}}^2\right), \quad (1)$$

which is a radial basis similarity based on a non-euclidean metric. This function has already been used by Chen et al. (2006) as an embedding for MIL. However, they have missed its importance and implications for sequential decision making.

3.3. Landmarks, earliness and learning formulation

As mentioned above, the decision function f is learned in a landmarking space defined by ψ. Such a choice raises the question of how to select the discriminative frames $\{\boldsymbol{p_j}\}_{j=1}^m$. A natural way to circumvent this issue is to select the relevant landmarks (during training) among all the frames available in the training dataset (Chen et al., 2006; Kar & Jain, 2012). This can be easily achieved thanks to an ℓ_1-penalization on the weight vector $\mathbf{w}$ (Tibshirani, 1996).

Interestingly, this landmark selection idea can be extended in order to promote earliness in the decision. By strongly penalizing the selection of late-appearing frames in fully observed sequences, we incite the decision function to compare a sequence $\mathbf{X}_{1..t}$ to early discriminative frames. This promotes earliness in the decision. Thus, we replace the ℓ_1-norm by a weighted norm $\|\mathbf{w}\|_{\ell_1}^{\boldsymbol{\mu}} = \sum_{j=1}^m \mu_j |\mathrm{w}_j|$. Here, $\boldsymbol{\mu}$ is a predefined weighting vector, the components of which are small for early-appearing landmarks (typically 1) and progressively larger for later ones.

The learning problem of the detector $f = \langle \mathbf{w} \mid \psi(\cdot) \rangle_{\ell_2}$ (jointly with the detection threshold b) is then obtained by writing down an ℓ_1-norm SVM (Zhu et al., 2004) with the features mentioned above (that is, the positivity constraint for reliability and the weighted norm for earliness):

$$\begin{array}{ll} \underset{\mathbf{w},\, b,\, \boldsymbol{\xi}}{\text{minimize}} & \|\mathbf{w}\|_{\ell_1}^{\boldsymbol{\mu}} + C \sum_{i=1}^n \xi_i \\ \text{s. t.} & \left\{ \begin{array}{l} y_i \left(\left\langle \mathbf{w} \mid \psi(\mathbf{X}_{1..T}^{(i)}) \right\rangle_{\ell_2} - b \right) \geq 1 - \xi_i, \forall i \\ \boldsymbol{\xi} \succcurlyeq 0,\ \mathbf{w} \succcurlyeq 0, \end{array} \right. \end{array} \quad (2)$$

where C is a positive trade-off parameter. Problem (2) is a linear program that can be solved using off-the-shelf tools such as *lpsolve* (Berkelaar et al., 2004). The only potential embarrassment could be the size of $\mathbf{w}$, as $\mathbf{w} \in \mathbb{R}^{nT}$, resulting from considering all the frames available in the training dataset as landmarks. This point is tackled in Section 5.

Remark Despite what has been said before, it is quite important to understand that in some situations, earliness can not be controlled. This is so if the discriminative frames are not expected to appear in a structured manner. For instance if their probability of appearance is equally distributed over the time-frame $[1, T]$. A concrete example is a shout in an audio recording of a kid play area. In this case, we face what we call a non-structured event. Thus, playing with $\boldsymbol{\mu}$ may be harmful and going back to a usual ℓ_1-norm could be in our best interest.

Remark An alternative way used in the literature for promoting early detection, is to make use of partially-observed sequences in the learning procedure, through an augmented loss function (Ellis et al., 2013; Hoai & De la Torre, 2014). By forcing them to be well detected, it is possible to control the earliness of the decision. However it usually leads to complex optimization problems which are slow to solve as increasing score constraints such as $f(\mathbf{X}_{1..t_1}) \leq f(\mathbf{X}_{1..t_2})$, for $t_1 \leq t_2$, are explicitly stated in the learning problem.

3.4. Generalization guarantee

Proposition (3.1) claims that our detector f is reliable but the proof gives another interesting result: $g(\mathbf{X}_{1..T}) = \max_{1\leq t\leq T} f(\mathbf{X}_{1..t}) = f(\mathbf{X}_{1..T})$. In other words, the decision function in the sequential prediction phase g is identical to the learned one f. This accounts for our interest in guaranteeing the generalization capability of our detector. Any result on the learned (over the full sequences) function

f is still true on the function g used in practice.

In the proposed framework, the detector f is trained by minimizing the empirical loss $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(f(\mathbf{X}_{1..T}^{(i)}), y_i\right)$ over a function class $\mathcal{F}$. The equivalence with Problem (2) can be obtained thanks to the theory by Tikhonov & Arsenin (1977). Here, the loss function is $\ell(z,y) = \max(0, 1 - y(z-b))$ and $\mathcal{F} = \left\{\mathbf{x} \mapsto \langle \mathbf{w} \mid \mathbf{x}\rangle_{\ell_2} : \mathbf{w} \in \mathbb{R}^m, \mathbf{w} \succcurlyeq 0, \|\mathbf{w}\|_{\ell_1}^{\boldsymbol{\mu}} \leq c_1\right\}$.

This point of view enables us to give a theoretical support to every early detector learned through the proposed framework. First, let us remind the definition given by Kakade et al. (2009) of the Rademacher complexity of a function class $\mathcal{F}$: $\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}_{1..T}^{(i)})\epsilon_i\right]$, where $\epsilon_1, \ldots, \epsilon_n$ are random variables, that independently take values in $\{-1, +1\}$ with equal probability. As usual, the sequences from the training dataset are supposed independent and identically distributed. The following proposition claims that the complexity of our class of early detector is bounded in $O(n^{-\frac{1}{2}})$.

Proposition 3.2 *Suppose that* $\exists c_\infty \geq 0 \colon |k(\mathbf{X}_{1..t}, \boldsymbol{p})| \leq c_\infty$ *for all* $(\boldsymbol{p}, \mathbf{X}_{1..t}) \in \mathcal{S} \times \mathcal{X}_T$. *If the lowest component of* $\boldsymbol{\mu}$ *is* 1*, then:*

$$\mathcal{R}_n(\mathcal{F}) \leq c_1 c_\infty \sqrt{\frac{2\log m}{n}}.$$

Proof Direct application of (Kakade et al., 2009, Theorem 3, Example (2)) in the landmarking space ($\mathbf{x} = \psi(\mathbf{X}_{1..t}) \in \mathbb{R}^m$), considering that the weights of $\|\cdot\|_{\ell_1}^{\boldsymbol{\mu}}$ are greater than 1. ■

This proposition can be used to derive a generalization bound thanks to (Bartlett & Mendelson, 2002; Kakade et al., 2009, theorem 1). This one states that, with high probability, the expected loss $L(g) = \mathbb{E}\left[\ell\left(g(\mathbf{X}_{1..T}), y\right)\right]$ is uniformly bounded in the following way:

$$L(g) \leq \hat{L}(f) + 2\mathcal{R}_n(\mathcal{F}) + O(n^{-\frac{1}{2}}) = \hat{L}(f) + O(n^{-\frac{1}{2}}).$$

This can be interpreted as: the expected *real-time* loss tends to be low if the empirical loss (minimized during the learning procedure) is low too and/or if the sample of training sequences grows.

4. Relaxing reliability constraints

Reliability is highly desirable for early detection although few approaches in the literature satisfy this property. Parrish et al. (2013) provide some probabilistic property while the main method we compare our framework with, called MMED (Hoai & De la Torre, 2014), is not reliable.

In our framework, we induce a deterministic reliability property by imposing some positivity constraints on the decision function weights. However, such constraints tend to reduce the expressivity of the decision function leading to poorer capability of learning the training examples. Hence, we propose to learn our decision function $f(\cdot)$ by making a compromise between reliability and expressivity. This trade-off is induced by relaxing the positivity constraints, yielding the following learning problem:

$$\begin{array}{ll} \underset{\mathbf{w},\, b,\, \boldsymbol{\xi},\, \boldsymbol{v}}{\text{minimize}} & \|\mathbf{w}\|_{\ell_1}^{\boldsymbol{\mu}} + C\sum_{i=1}^{n}\xi_i + \lambda\sum_{j=1}^{m} v_j \\ \text{s.t.} & \begin{cases} y_i\left(\left\langle \mathbf{w} \mid \psi(\mathbf{X}_{1..T}^{(i)})\right\rangle_{\ell_2} - b\right) \geq 1 - \xi_i, \forall i \\ \boldsymbol{\xi} \succcurlyeq 0, \mathbf{w} \succcurlyeq -\boldsymbol{v}, \boldsymbol{v} \succcurlyeq 0, \end{cases} \end{array} \tag{3}$$

where λ is a positive trade-off parameter between expressivity and reliability. Note that when $\lambda \to +\infty$, we get back the original Problem (2) while when $\lambda = 0$, we have a classical learning problem similar to the one proposed in (Chen et al., 2006) but with an earliness-promoting penalty.

Because of this relaxation, Proposition 3.1 guaranteeing reliability does not hold anymore. Interestingly, the next proposition tells us that we can easily derive for the learned $f(\cdot)$ a reliable decision function.

Proposition 4.1 *Let* $k \colon \mathcal{X}_T \times \mathcal{S} \to (-\infty, 1]$ *be a non-decreasing time-dependent function. For any model* $f = \langle \mathbf{w} \mid \psi(\cdot)\rangle_{\ell_2}$*, let* $\mathcal{P} = \{j \in \{1, \ldots, m\} : w_j > 0\}$ *and* $\mathcal{N}$ *the sets indexing respectively positive and negative weights. Then* $\hat{f}(\cdot) \triangleq \sum_{j\in\mathcal{P}} \mathrm{w}_j k(\cdot, \boldsymbol{p_j}) + \sum_{j\in\mathcal{N}} \mathrm{w}_j$ *is reliable with respect to any detection threshold* b*. Moreover, if* $\hat{f}(\mathbf{X}_{1..t}) \geq b$*, then* $f(\mathbf{X}_{1..T}) \geq b$.

Proof First, reliability comes from the same argument as in Proposition 3.1. Second, $f(\mathbf{X}_{1..T}) - \hat{f}(\mathbf{X}_{1..t}) = \sum_{j\in\mathcal{P}} \mathrm{w}_j(k(\mathbf{X}_{1..T}, \boldsymbol{p_j}) - k(\mathbf{X}_{1..t}, \boldsymbol{p_j})) + \sum_{j\in\mathcal{N}} \mathrm{w}_j(k(\mathbf{X}_{1..T}, \boldsymbol{p_j}) - 1) \geq 0$ since k is non-decreasing with respect to t and $k(\mathbf{X}_{1..T}, \boldsymbol{p_j}) \leq 1$. Thus, $f(\mathbf{X}_{1..T}) \geq b$ as soon as $\hat{f}(\mathbf{X}_{1..t}) \geq b$. ■

Several remarks can be stated from this proposition. First, one important statement is that any model can now achieve reliable decision. This ability of delivering reliable detection at a time t depends on how negative is the score $\sum_{j\in\mathcal{N}} \mathrm{w}_j$. This latter score can be understood as a gap needed for a relaxed model for ensuring reliability. Hence, if $\mathrm{w}_j = 0$ for all $j \in \mathcal{N}$ without specifying explicitly this constraint in the learning problem, then we obtain the same reliability condition than the one given by Proposition 3.1.

The second important remark is that since $\sum_{j\in\mathcal{P}} \mathrm{w}_j k(\mathbf{X}_{1..t}, \boldsymbol{p_j})$ is a non-decreasing function, our relaxed model actually trades earliness versus reliability. Indeed, at a time t_1, we can have $f(\mathbf{X}_{1..t_1}) \geq b$

without reliability but we need to wait until $t_2 > t_1$ to get $\hat{f}(\mathbf{X}_{1..t_2}) \geq b$.

5. Training algorithm

Like Problem (2), Problem (3) can be solved with an off-the-shelf solver, by using the decoupling trick: $\|\mathbf{w}\|_{\ell_1}^{\boldsymbol{\mu}} = \sum_{j=1}^{m} \mu_j(\mathrm{w}_j^+ + \mathrm{w}_j^-)$, with $\mathbf{w} = \mathbf{w}^+ - \mathbf{w}^-$ and $\mathbf{w}^+, \mathbf{w}^- \succcurlyeq 0$. However, to do so, one has to accept to deal with many components of $\mathbf{w}$ that will be null at the optimality. In order to speed up the training of our detector, we propose to solve the dual of Problem (3) instead:

$$\begin{array}{ll} \underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\text{maximize}} & \langle \mathbb{1} \mid \boldsymbol{\beta} \rangle_{\ell_2} \\ \text{s. t.} & \left\{ \begin{array}{l} 0 \preccurlyeq \boldsymbol{\beta} \preccurlyeq C\mathbb{1} \\ -\boldsymbol{\mu} - \lambda\mathbb{1} \preccurlyeq \boldsymbol{Q}\boldsymbol{\beta} \preccurlyeq \boldsymbol{\mu} \\ \langle \boldsymbol{\beta} \mid \boldsymbol{y} \rangle_{\ell_2} = 0, \end{array} \right. \end{array} \tag{4}$$

where $\boldsymbol{Q} = \left(y_1\psi(\mathbf{X^{(1)}}_{1..T}, \ldots, y_1\psi(\mathbf{X^{(n)}}_{1..T})\right) \in \mathbb{R}^{m\times n}$. The primal variables $\mathbf{w}$ and b turn out to be the Lagrangian variables of the last two constraints in (4). Moreover, Karush-Kuhn-Tucker conditions indicate that if $\mu_j - \lambda < (\boldsymbol{Q}\boldsymbol{\beta})_j < \mu_j$, then $\mathrm{w}_j = 0$. This suggests a column generation algorithm (Nocedal & Wright, 2000). Such a procedure is presented in Algorithm 1. Convergence is guaranteed by convex optimization theory (Luenberger, 1984).

Algorithm 1 Algorithm to learn a an early detector.

$\mathcal{A} \leftarrow$ random sample of indexes between 1 and m
repeat
  $\boldsymbol{Q} \leftarrow \left(y_i k(\mathbf{X}^{(i)}_{1..T}, \boldsymbol{p_j})\right)_{j\in\mathcal{A}, 1\leq i\leq n}$
  $\boldsymbol{\beta} \leftarrow$ solve Problem (4)
  $j^- \leftarrow \underset{1\leq j\leq m}{\arg\min}\ ((\boldsymbol{Q}\boldsymbol{\beta})_j + \mu_j)$
  $j^+ \leftarrow \underset{1\leq j\leq m}{\arg\max}\ ((\boldsymbol{Q}\boldsymbol{\beta})_j - \mu_j)$
  if $-\mu_{j^-} - \lambda \leq (\boldsymbol{Q}\boldsymbol{\beta})_{j^-}$ and $(\boldsymbol{Q}\boldsymbol{\beta})_{j^+} \leq \mu_{j^+}$ **then**
    convergence
  else if $-\mu_{j^-} - \lambda > (\boldsymbol{Q}\boldsymbol{\beta})_{j^-}$ **then**
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{j^-\}$
  else
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{j^+\}$
  end if
until convergence

6. Numerical experiments

Our detection task is a multi-objective problem, where them goal is to correctly detect all the events with the fewest false alarm as possible, and in an early and reliable way. As such, it is difficult, if not impossible, to optimize both earliness and accuracy while being reliable. Hence, depending on the parameter $\boldsymbol{\mu}$ (promoting earliness) and the hyper-parameter λ controlling reliability, we can achieve models that perform well on a criterion but poorly on the other. We thus believe that the choice of which criteria should be put forward and thus the model selection is application and user dependent. Hence, in the sequel, we have presented the results obtained by several of our models. As competitors, we have considered MMED, which is not a reliable detector and the MILES classifier by Chen et al. (2006), that we turned into a reliable sequential detector. Note that although we present the *sequential and reliable MILES* as a competitor, this algorithm is a contribution of ours as it is subsumed under our model SimpleED ($\boldsymbol{\mu} = \mathbb{1}$ and $\lambda = 0$) and strongly relies on Proposition 4.1.

In the whole section, AUROC refers to the area under the receiver operating curve obtained for $\hat{g}(\mathbf{X}_{1..T}) = \max_{1\leq t\leq T} \hat{f}(\mathbf{X}_{1..t})$ (that is for a sequential test). It measures the overall capability of detection independently of the threshold b (1 notifies a perfect ability). Following (Hoai & De la Torre, 2014), we also consider the activity monitoring operating curve (AMOC) (Fawcett & Provost, 1999), which depicts the average normalized-time-to-detect the occurrences of the event versus the false positive rate. This curve is obtained by making the detection threshold b vary. To perform a fair comparison, independently of the trade-off between accuracy and earliness, we analyze the area under the AMOC curve (denoted AUAMOC). Unless specified, numerical results presented (AUROC, AUAMOC, AMOC curve, training time) are averaged on 10 random runs, where the models are evaluated on a test dataset, after being learned on a separate training dataset. The parameter γ that defines the landmarking space is set to 2^{-1} (default value for normalized data). Smallest value of $\boldsymbol{\mu}$ is always set to 1 according to Proposition 3.2. The other values are defined following a linear trend. C is obtained through a 5-fold cross-validation (maximizing the AUROC) on the following grid $[2^{-2}, 2^0, \ldots, 2^{10}]$. Eventually, MMED is trained in accordance to its design, our framework and the other confronted methods, using the time-serie $\{\psi(\mathbf{X}_{1..t})\}_{t=1}^{T}$: event occurrences are sent with the time frame label $[1, T]$ (meaning that the whole sequence is an occurrence) and the other sequences with the time frame label $[0, 0]$ (meaning that there is no occurrence in this sequence). A quick comparison reported a faster training and slightly better results than drowning the occurrences in large time-series. Matlab® code ran on a single core of an Intel® Xeon® E5-2630 CPU, operating at 2.4 GHz with GNU/Linux and 144 Gb of RAM. In addition, this code is available on the authors' websites.

MODEL			CRITERION		
NUMBER	λ	max($\boldsymbol{\mu}$)	AUAMOC	AUROC	TIME (S)
1	0	1	0.8309	1.000	0.27
2	0	2	0.1961	0.998	0.28
3	1	2	0.5425	1.000	0.30
4	2	2	0.5590	1.000	0.29
5	INF	2	0.5601	1.000	0.25
6	0	4	0.0630	0.995	0.27
7	1	4	0.0797	0.998	0.27
8	2	4	0.1313	1.000	0.28
9	INF	4	0.2435	1.000	0.28
10	0	8	0.0587	0.995	0.26
11	1	8	0.0657	0.997	0.27
12	2	8	0.0786	0.998	0.28
13	INF	8	0.0986	0.999	0.23
MMED	-	-	0.1066	1.000	32.93

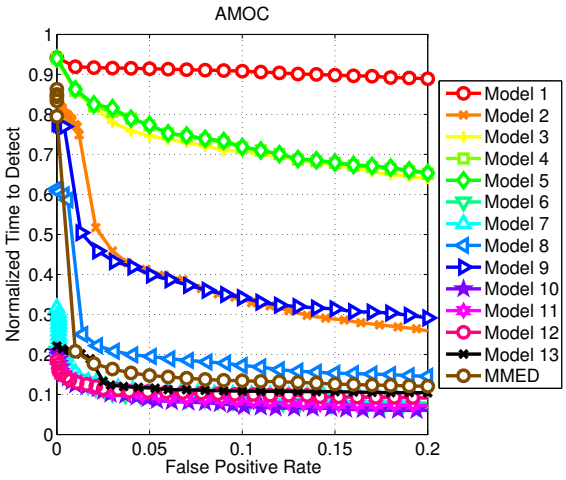


Figure 1. Toy dataset: table summarizing numerical results (left) and AMOC curves of the different models (right).

6.1. Toy dataset

This first numerical experiment is aimed at assessing the ability of SimpleED to promote an early detection. This experiment is performed with a toy dataset, which is made up of two classes. Each class is a one-second linear chirp (both starting at 100 Hz and ending at respectively 7000 and 8000 Hz) with an additive Gaussian noise. We use MFCC computed on a sliding window as time-feature representation. The dataset contains 400 sequences of 40 frames. For each run, half of them are selected for training, and the other half is for testing the models. In order to make MMED tractable, we have pruned the number of landmarks by a factor of 2, yielding thus 4000 landmarks. Figure 1 depicts the AMOC curves for SimpleED under different parameters μ and λ. The first model (Model 1) with the slope of μ equals to 1 and $\lambda = 0$ corresponds to the reliable MILES model by Chen et al. (2006) which as been made reliable according to Proposition (4.1). Performance of MMED is also presented. From the table on Figure 1, several interesting remarks can be pointed out. First, MILES (Model 1) performs poorly in term of AUAMOC. This a natural consequence of not imposing any constraints on the landmarks. When keeping fixed the penalty λ on the negative weights (encouraging reliability), increasing the largest weight on the ℓ_1-penalty, reduces AUAMOC. For $\lambda = 2$, AUAMOC takes values 0.56, 0.13 and 0.08 respectively for largest values of μ (2, 4 and 8). This is a natural consequence of imposing early landmarks to be selected. For similar weights on the ℓ_1-penalty, inducing reliability increases AUAMOC and AUROC (see for instance the results for $\mu = 4$). We can explain this by two points: i) reliability induces model to select landmarks for which it has more confidence (hence later landmarks); ii) AUROC increases because models which are less constrained on reliability (smaller values of λ) but are used in a reliable decision context, have larger negative weights that need to be compensated before detection. Hence these models tend to miss some positive time-series (more miss to detect).

We can note that MMED is on par with SimpleED in terms of AUROC and slightly worse for AUAMOC. Moreover, its normalized-time to detect is poorer than those of most of SimpleED for small false positive rate, which is the interesting setting. In addition, its running time for training is almost 100 times more expensive than SimpleED.

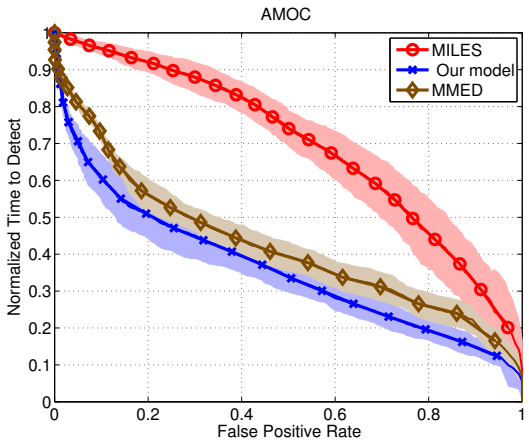
6.2. BCI data

Here, we address electroencephalographic event detection related to brain-computer interfaces (BCI). The objective is to detect 5-second length event occurrences. In BCI tasks, being able to reliably detect an event occurrence as soon as possible is of primary importance, since it helps in increasing command bitrate related to the BCI. Again, reliability is a key feature as it is preferable to collect more data than to make an error due to incomplete observation. Errors are indeed difficult to fix with BCI.

For this experiment, we used the publicly available dataset 2a from BCI competition IV (Brunner et al., 2008). The features we extracted for a single frame are similar to the classical procedure used in BCI and they are based on Common Spatial Patterns (CSP) (Lotte & Guan, 2011). The main difference here is that instead of projecting the whole 5s signals onto the CSP subspace, frames of 50-ms window are built and the portion of related signals are projected. We thus have 100 frame per signals, 144 signals per subject and 9 subjects. Again, in order to make MMED tractable, we have pruned the number of landmarks by a factor of 3, yielding thus 4800 landmarks. Because the BCI frames are not normalized, the parameter γ used for computing similarity in Equation (1) has been set to the inverse

Table 1. BCI: statistics for a subset of our models (due to space limitation). Model 1 is *sequential* MILES. The penultimate line provides a single average criterion for comparing the methods (the larger the better).

SUBJECT	CRITERION	MODEL						
		1	4	5	8	9	13	MMED
1	AUAMOC	0.56	0.56	0.52	0.61	0.56	0.56	0.26
	AUROC	0.83	0.89	0.94	0.89	0.95	0.93	0.93
2	AUAMOC	0.79	0.80	0.81	0.79	0.80	0.81	0.66
	AUROC	0.54	0.53	0.53	0.52	0.53	0.52	0.49
3	AUAMOC	0.54	0.55	0.47	0.48	0.48	0.46	0.29
	AUROC	0.84	0.92	0.92	0.92	0.92	0.94	0.95
4	AUAMOC	1.00	0.86	0.88	0.85	0.88	0.88	0.77
	AUROC	0.50	0.55	0.60	0.51	0.57	0.58	0.57
5	AUAMOC	0.73	0.78	0.79	0.79	0.85	0.89	0.72
	AUROC	0.50	0.50	0.56	0.46	0.55	0.52	0.56
6	AUAMOC	0.74	0.81	0.74	0.75	0.73	0.77	0.53
	AUROC	0.57	0.61	0.65	0.59	0.63	0.65	0.69
7	AUAMOC	0.78	0.68	0.86	0.73	0.86	0.72	0.59
	AUROC	0.67	0.79	0.62	0.71	0.61	0.74	0.70
8	AUAMOC	0.36	0.37	0.37	0.37	0.37	0.39	0.12
	AUROC	0.95	0.94	0.94	0.93	0.93	0.94	0.99
9	AUAMOC	0.18	0.19	0.19	0.25	0.26	0.19	0.24
	AUROC	0.97	0.97	0.97	0.98	0.97	0.97	0.96
(1-AUAMOC +AUROC)/2		0.54	0.56	0.56	0.55	0.55	0.56	0.65
	TIME (S)	1.72	1.59	1.47	1.48	1.36	1.25	34.09



MODEL	AUAMOC	AUROC	(1-AUAMOC +AUROC)/2	TIME (S)
MILES	0.69	0.53	0.42	0.122
SIMPLEED	0.36	0.88	0.76	0.096
MMED	0.42	0.93	0.75	2.969

Figure 2. Emotions dataset: AMOC curves with standard deviations (shaded areas) and table providing averaged statistics.

of the square-root of the average distance between 20% of the frames.

Results are presented in Table 1. They are single run results since for each subject training and testing sets are pre-defined. They clearly show that SimpleED, whose parameters are identical to the toy problem, performs slightly worse than MMED regarding the average AMOC and AUROC over all the subjects. We understand this as the little price to pay for reliability! This can be clearly seen for AUROC-best performing subjects (1, 3, 8 and 9), where earliness is traded against reliability. Again, our algorithm achieved a large gain in training time compared to MMED.

6.3. Emotions dataset

As a video-based experiment, we consider the extended Cohn-Kanade dataset (CK+) (Lucey et al., 2010). This one contains 327 facial image sequences from 123 subjects. Each subject performed a prescribed emotion (among anger, contempt, disgust, fear, happiness, sadness, and surprise) from neutral behavior (first frame) to peak expression (last frame). Similarly to (Hoai & De la Torre, 2014), we consider the task of detecting negative emotions: anger, disgust, fear, and sadness. As a time-feature representation, we use the tracking landmarks of the Active Appearance Models, which are publicly available (Lucey et al., 2010), normalized by subtracting the features of the first and neutral frame. We used 100 time-series for training leading to 1879 landmarks.

Comparison results are given in Figure 2. For this experiment, we have kept the general experimental set-up but in addition, we have cross-validated the parameter λ among the values $(0, 2^{-5}, 2^{-3}, 2^{0}, 2^{3}, 2^{5}, \infty)$. From the table, we can see that MMED and SimpleED are on par, with an advantage for SimpleED in AUAMOC. Here, again the training running time of our algorithm is of an order of magnitude better than the one of MMED. From the figure, we notice that the AMOC curve of SimpleED is always better than the one of MMED and there exists large gap of normalized-time-to-detect around 0.1 false positive rate in our favor. This again shows that the approaches we developed are efficient as the state-of-the-art for early detection, while being in addition reliable.

7. Conclusion

In this paper, we have provided a novel framework for early and reliable detection of temporal events. The detector is built upon a landmarking space with specific properties. We have also investigated how these properties can be relaxed in order to enhance the expressivity power of the decision function and provide novel ways for guaranteeing reliability of these models. Experimental results highlight that our detector based on similarity functions is faster to train and achieves similar performances than its competitor MMED while providing reliable decisions and being far more efficient to learn.

Constraints relaxations have been considered for performance boosting. Another way to achieve this goal would be to integrate a representation learning into the global learning strategy. Learning the landmarks would also be an important step in the future improvements of this framework.

Acknowledgments

This work was partially funded by the industrial chair "Machine Learning for Big Data", CEA, LIST and the *Direction Générale de l'Armement* (French Ministry of Defense).

References

Bartlett, P.L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Berkelaar, M., Eikland, K., and Notebaert, P. lpsolve: Open source (mixed-integer) linear programming system. Eindhoven University of Technology, 2004.

Brunner, C., Leeb, R., Mller-Putz, G.R., Schlgl, A., and Pfurtscheller, G. {BCI} {Competition} 2008 – {Graz} data set {A}. Technical report, Institute for Knowledge Discovery, Institute for Human-Computer Interfaces, Graz University of Technology, Austria, 2008. URL `http://www.bbci.de/competition/iv/`.

Chen, Y., Bi, J., and Wang, J.Z. Miles: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1931–1947, 2006.

Dietterich, T.G., Lathrop, R.H., and Lozano-Prez, T. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

Ellis, C., Masood, S.Z., Tappen, M.F., Laviola, J.J. Jr., and Sukthankar, R. Exploring the trade-off between accuracy and observational latency in action recognition. *International Journal of Computer Vision*, 101:420–436, 2013. ISSN 0920-5691.

Fawcett, T. and Provost, F. Activity monitoring: Noticing interesting changes in behavior. In *SIGKDD conference on knowledge discovery and data mining*, 1999.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29: 2247–2253, 2007.

Hoai, M. and De la Torre, F. Max-margin early event detectors. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Hoai, M. and De la Torre, F. Max-margin early event detectors. *International Journal of Computer Vision*, 107: 191–202, 2014.

Kakade, S.M., Sridharan, K., and Tewari, A. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems*. 2009.

Kar, P. and Jain, P. Supervised learning with similarity functions. In *Advances in Neural Information Processing Systems*. 2012.

Keeler, J.D., Rumelhart, D.E., and Leow, W.-K. Integrated Segmentation and Recognition of Hand-Printed Numerals. In *Advances in Neural Information Processing Systems*. 1991.

Lotte, F. and Guan, C. Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms. *IEEE Transactions on Biomedical Engineering*, 58(2):355–362, 2011.

Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 2010. doi: 10.1109/CVPRW.2010.5543262.

Luenberger, D.G. *Linear and nonlinear programming*. Addison-Wesley, 1984. ISBN 9780201157949.

Neill, D., Moore, A., and Cooper, G. A bayesian spatial scan statistic. In *Advances in Neural Information Processing Systems*. 2005.

Nocedal, J. and Wright, S.J. *Numerical optimization*. Springer, 2000. ISBN 9780387303031.

Parrish, N., Anderson, H.S., Gupta, M.R., and Hsiao, D.Y. Classifying with confidence from incomplete information. *Journal of Machine Learning Research*, 14:3561–3589, 2013.

Rodriguez, J.J. and Alonso, C.J. Boosting interval-based literals: Variable length and early classification. In *Workshop on Knowledge Discovery from (Spatio-) Temporal Data*, 2002.

Sangnier, M., Gauthier, J., and Rakotomamonjy, A. Early frame-based detection of acoustic scenes. In *IEEE International Workshop on Applications of Signal Processing to Audio and Acoustics*, 2015.

Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Tikhonov, A.N. and Arsenin, V.Y. *Solutions of ill-posed problems*. Winston, Washington, DC, 1977.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.

Wang, Jun and Zucker, J.-D. Solving the multiple-instance problem: A lazy learning approach. In *International Conference on Machine Learning*, 2000.

Xing, Z., Pei, J., and Yu, P.S. Early prediction on time series: A nearest neighbor approach. In *International Joint Conferences on Artificial Intelligence*, 2009.

Xing, Z., Pei, J., and Yu, P.S. Early classification on time series. *Knowledge and Information Systems*, 31:105–127, 2012. ISSN 0219-1377.

Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R. 1-norm Support Vector Machines. In *Advances in Neural Information Processing Systems*. 2004.