Label Noise-Tolerant Hidden Markov Models for Segmentation: Application to ECGs

Benoît Frénay, Gaël de Lannoy*, and Michel Verleysen

Machine Learning Group, ICTEAM Institute, Université catholique de Louvain
3 place du Levant, B-1348 Louvain-la-Neuve, Belgium

Abstract. The performance of traditional classification models can adversely be impacted by the presence of label noise in training observations. The pioneer work of Lawrence and Schölkopf tackled this issue in datasets with independent observations by incorporating a statistical noise model within the inference algorithm. In this paper, the specific case of label noise in non-independent observations is rather considered. For this purpose, a label noise-tolerant expectation-maximisation algorithm is proposed in the frame of hidden Markov models. Experiments are carried on both healthy and pathological electrocardiogram signals with distinct types of additional artificial label noise. Results show that the proposed label noise-tolerant inference algorithm can improve the segmentation performances in the presence of label noise.

Keywords: label noise, hidden Markov models, expectation maximisation algorithm, segmentation, electrocardiograms.

1 Introduction

In standard situations, supervised machine learning algorithms learn their parameters to fit previously labelled data, called training observations, as best as possible. In real situations, however, it is difficult to guarantee perfect labelling, e.g. because of the subjectivity of the labelling task, of the lack of information or of communication noise. In particular, label errors are likely to arise in biomedical applications involving the tedious and time-consuming labelling of a large amount of data by one or several medical experts. The label noise issue is typically addressed in regression problems by assuming independent Gaussian noise on the regression target. In classification problems, although standard algorithms such as support vector machines are able to cope with outliers and feature noise to some degree, the label noise issue is however mostly left untreated.

Previous work addressing the label noise issue incorporated a noise model into a generative model which assumes independent and identically distributed (i.i.d.) observations [1–3]. Nevertheless, this issue is mostly left untreated in the case of models for the segmentation of sequential (non i.i.d.) observations such as hidden Markov models (HMMs). In this work, a variant of HMMs which is

* Gaël de Lannoy is funded by a Belgian F.R.I.A. grant.

D. Gunopulos et al. (Eds.): ECML PKDD 2011, Part I, LNAI 6911, pp. 455–470, 2011.

robust to the label noise is proposed. To illustrate the relevance of the proposed model, artificial electrocardiogram (ECG) signals generated using ECGSYN [4] and real ECG recordings from the Physiobank database [5] are used in the experiments. The label noise issue is indeed known affect the segmentation of waveform boundaries by experts in ECG signals [6]. Nevertheless, the proposed model also applies to any kind of sequential data facing the label noise issue, for example biomedical signals such as EEGs, EMGs and many others.

This paper is organised as follows. Section 2 reviews related work. Section 3 introduces hidden Markov models and two standard inference algorithms. Section 4 derives a new, label noise-tolerant algorithm. Section 5 quickly reviews electrocardiogram signals and details the experimental settings. Finally, empirical results are presented in Section 6 and conclusions are drawn in Section 7.

2 Related Work

Before presenting the state-of-the-art in classification with label noise, it is first important to distinguish the label noise issue from the semi-supervised paradigm where some data points in the training set are completely left unlabelled. Here, we rather consider the framework where an unknown proportion of the observations are wrongly labelled. To our knowledge, existing approaches to this problem are relatively few. These approaches can be divided in three categories: filtering approaches, model-based approaches and plausibilistic approaches.

Filtering techniques act as a preprocessing of the training set to either remove noisy observations or correct their labels. These methods involve the use of a criterion to detect mislabelled observations. For example, [7] uses disagreement in ensemble methods. Furthermore, [8] introduces an algorithm to iteratively modify the examples whose class label disagrees with the class labels of most of their neighbours. Eventually, [9] uses information gain to detect noisy labels.

On the other hand, model-based approaches tackle the label noise by incorporating the mislabelling process as an integral part of the probabilistic model. Pioneer work by [1] incorporated a probabilistic noise model in a kernel-Fisher discriminant for binary classification. Later, [2] extended this model by relaxing the Gaussian distribution assumption and carried out extensive experiments on more complex datasets, which convincingly demonstrated the value of explicit label noise modeling. More recently the same model has been extended to multi-class datasets [10]. Bouveyron et al. also proposes a distinct robust mixture discriminant analysis [3], which consists in two steps: (i) learning an unsupervised Gaussian mixture model and (ii) computing the probability that each cluster belongs to a given class.

Eventually, plausibilistic approaches assume that the experts have explicitly provided uncertainties over labels. Specific algorithms are then developed to integrate and to focus on such uncertainties [11].

This work concentrates on model-based approaches to embed the noise process into classifiers. Model-based approaches have a sound theoretical foundation and tackle the noise issue in a more principled and transparent manner without

discarding potentially useful observations. Our contribution in this field is the development of a label noise-tolerant hidden Markov model for labelling of sequential (non i.i.d.) observations.

3 Hidden Markov Models for Segmentation

This section introduces hidden Markov models for segmentation. Two widely used inference algorithms are detailed: supervised learning and the Baum-Welch algorithm. Their application to ECG segmentation is discussed in Section 5.

3.1 Hidden Markov Models

HMMs are probabilistic models of time series generating processes where two distinct sequences are considered: the states $S_1 \dots S_T$ and observations $O_1 \dots O_T$. Here T is the length of these sequences (see Fig. 1). At a given time step t, the current observation O_t and the next state S_{t+1} are considered to depend only on the current state S_t. For example, in the case of ECGs, the process under study is the human heart. Hence, states and observations correspond to the inner state and electrical activity of the heart, respectively (see Section 5 for more details).

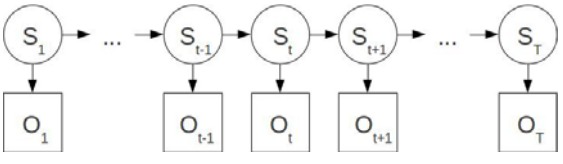

Fig. 1. Conditional dependencies in an hidden Markov model

Using the independence assumptions introduced above, an HMM is completely specified by its set of parameters $\Theta = (q, a, b)$ where q_i is the prior of state i, a_{ij} is the transition probability from state i to state j and b_i is the observation distributions for state i [12]. Usually, b_i is modelled by a Gaussian mixture model (GMM) with parameters $(\pi_{ik}, \mu_{ik}, \Sigma_{ik})$ where π_{ik}, μ_{ik} and Σ_{ik} are the prior, mean and covariance matrix of the kth Gaussian component, respectively.

Given a sequence of observations with expert annotations, the HMM inference problem consists in learning the parameters Θ from data. The remaining of this section presents two approaches for estimating the parameters. Once that an HMM is inferred, the segmentation of new signals can be done using the Viterbi algorithm, which looks for the most probable state sequence [12].

3.2 Algorithms for Inference

A simple solution to infer an HMM from data consists in assuming that the expert annotations are correct and trustworthy. Given this assumption, q and a are simply obtained by counting the state occurrences and transitions in the

data. Then, each observation distribution is fitted using the observations labelled accordingly. This approach has the advantage of being simple to implement and having a very low computational cost. However, if the labels are not perfect and polluted by some label noise, the produced HMM may be significantly altered.

The Baum-Welch algorithm is another, unsupervised algorithm [12]. More precisely, it assumes that the true labels are unknown, i.e. it ignores the expert annotations. The likelihood of the observations is maximised using an expectation-maximisation (EM) scheme [13], since no closed-form maximum likelihood estimator is available in this case. During the E step, the posteriors $P(S_t = i|O_1 \dots O_T)$ and $P(S_{t-1} = i, S_t = j|O_1 \dots O_T)$ are estimated for each time step t and states i and j. Then, these posteriors are used during the M step in order to estimate the prior vector q, the transition matrix a and the observation distributions b_i.

The main advantage of Baum-Welch is that wrong expert annotations should have no impact on the inferred HMM. However, in practice, expert annotations are used to compute a initial estimate of the HMM parameters, which is necessary for the first E step. Moreover, ignoring expert annotations can also be a disadvantage: if the expert uses a specific decomposition of the ECG dynamic, such subtleties may be lost in the unsupervised learning process.

4 A Label Noise-Tolerant Algorithm

Two algorithms for HMM inference have been introduced in Section 3. However, neither of them is satisfying when label noise is introduced. On the one hand, supervised learning is bound to trust blindly the expert annotations. Therefore, as shown in Section 6, label noise can degrade the segmentation quality. On the other hand, the Baum-Welch algorithm fails to encode precisely the expert knowledge. Indeed, as shown experimentally in Section 6, even predictions on clean, easy-to-segment signals do not match accurately the expert annotations.

This section introduces a new algorithm for HMM inference which lies in-between supervised learning and the Baum-Welch algorithm: expert annotations are used, but the label noise is modelled during the inference process in order to decrease the influence of wrong annotations.

4.1 Label Noise Modelling

Previous works showed the value of explicit label noise modelling for i.i.d. data in classification [1–3]. Here, a similar approach is used for non-independent sequential data. Two distinct, yet related sequences of states are considered (see Fig. 2): the sequence of observed, noisy annotations Y and the sequence of hidden, true labels S. In this paper, Y_t is assumed to depend only on S_t, i.e. Y_t is a (possibly noisy) copy of S_t.

An additional quantity $d_{ij} = P(Y_t = j|S_t = i, \Theta)$ is introduced for each pair of states (i, j), which is called the annotation probability. In order to avoid overfitting, the annotations probabilities take in this paper the restricted form

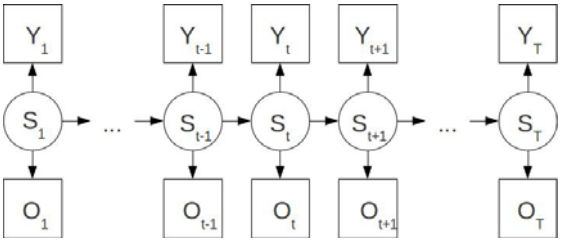

Fig. 2. Conditional dependencies in a label noise-tolerant hidden Markov model

$$d_{ij} = \begin{cases} 1 - p_i & (i = j) \\ \frac{p_i}{|\mathcal{S}|-1} & (i \neq j) \end{cases} \tag{1}$$

where p_i is the probability that the expert makes an error in state i and $|\mathcal{S}|$ is the number of possible states. Hence $d_{ii} = 1 - p_i$ is the probability of correct annotation in state i. Notice that d_{ij} is only used during inference. Here, Y is an extra layer put on a standard HMM to model the label noise. For segmentation, only the parameters linked to S and O are used, i.e. q, a, π, μ and Σ.

4.2 Finding the HMM Parameters with a Label Noise Model

Finding an estimate for both the HMM and label noise model parameters is achieved by maximising the incomplete log-likelihood

$$\log P(O, Y|\Theta) = \log \sum_S P(O, Y, S|\Theta), \tag{2}$$

where the sum spans all possible sequences of true states. As a closed-form solution does not exist, one can use the EM algorithm which is derived in the rest of this section. Notice that only approximate solutions are obtained, for EM algorithms are iterative procedures and may converge to local minima [13].

Definition of the $Q(\Theta, \Theta^{old})$ Function. The EM algorithm builds successive approximations of the incomplete log-likelihood in (2) and use them to maintain an estimate of the parameters [12, 14]. In the settings introduced above, it consists in alternatively (i) estimating the functional

$$Q(\Theta, \Theta^{old}) = \sum_S P(S|O, Y, \Theta^{old}) \log P(O, Y, S|\Theta) \tag{3}$$

using the current estimate Θ^{old} (E step) and (ii) maximising $Q(\Theta, \Theta^{old})$ with respect to the parameters Θ in order to update their estimate (M step). Since

$$P(O, Y, S|\Theta) = q_{s_1} \prod_{t=2}^{T} a_{s_{t-1}s_t} \prod_{t=1}^{T} b_{s_t}(o_t) \prod_{t=1}^{T} d_{s_t y_t}, \tag{4}$$

where o_t, y_t, s_1, s_{t-1} and s_t are the actual values taken by the random variables O_t, Y_t, S_1, S_{t-1} and S_t, the expression of $Q(\Theta, \Theta^{old})$ becomes

$$\sum_{i=1}^{|\mathcal{S}|} \gamma_1(i) \log q_i + \sum_{t=2}^{T} \sum_{i=1}^{|\mathcal{S}|} \sum_{j=1}^{|\mathcal{S}|} \epsilon_t(i,j) \log a_{ij} + \sum_{t=1}^{T} \sum_{i=1}^{|\mathcal{S}|} \gamma_t(i) \log b_i(o_t) + \sum_{t=1}^{T} \sum_{i=1}^{|\mathcal{S}|} \gamma_t(i) \log d_{iy_t} \quad (5)$$

where the posterior probabilities γ and ϵ are defined as

$$\gamma_t(i) = P(S_t = i | O, Y, \Theta^{old}) \quad (6)$$

and

$$\epsilon_t(i,j) = P(S_{t-1} = i, S_t = j | O, Y, \Theta^{old}). \quad (7)$$

E Step. The γ and ϵ variables must be computed in order to evaluate (5), which is necessary for the M step. In standard HMMs, these quantities are estimated during the E step by the forward-backward algorithm [12, 14]. Indeed, if forward variables α, backward variables β and scaling coefficients c are defined as

$$\alpha_t(i) = P(S_t = i | O_{1\ldots t}, Y_{1\ldots t}, \Theta^{old}) \quad (8)$$

$$\beta_t(i) = \frac{P(O_{t+1\ldots T}, Y_{t+1\ldots T} | S_t = i, \Theta^{old})}{P(O_{t+1\ldots T}, Y_{t+1\ldots T} | O_{1\ldots t}, Y_{1\ldots t}, \Theta^{old})} \quad (9)$$

$$c_t = P(O_t, Y_t | O_{1\ldots t-1}, Y_{1\ldots t-1}, \Theta^{old}), \quad (10)$$

one eventually obtains

$$\gamma_t(i) = \alpha_t(i) \beta_t(i) \quad (11)$$

and

$$\epsilon_t(i,j) = \alpha_{t-1}(i) c_t^{-1} a_{ij} b_j(o_t) d_{jy_t} \beta_t(j). \quad (12)$$

Here, the scaling coefficients c_t are introduced in order to avoid numerical issues. Indeed, for sufficiently large T (i.e. 10 or more), the dynamic range of both α and β will exceed the precision range of any machine. The scaling factors c_t are therefore introduced to keep the values within reasonable bounds [12]. The incomplete likelihood can be computed using $P(O, Y | \Theta^{old}) = \prod_{t=1}^{T} c_t$.

The forward-backward algorithm consists in using the recursive relationship

$$\alpha_t(i) c_t = \begin{cases} q_i b_i(o_1) d_{iy_1} & (t = 1) \\ b_i(o_t) d_{iy_t} \sum_{j=1}^{|\mathcal{S}|} a_{ji} \alpha_{t-1}(j) & (t > 1). \end{cases} \quad (13)$$

linking the α and c variables and the recursive relationship

$$\beta_t(i) = \begin{cases} 1 & (t = T) \\ \frac{1}{c_{t+1}} \sum_{j=1}^{|\mathcal{S}|} a_{ij} b_j(o_{t+1}) d_{jy_{t+1}} \beta_{t+1}(j) & (t < T) \end{cases} \quad (14)$$

linking the β and c variables. The scaling coefficients can be computed using the constraint $\sum_{i=1}^{|\mathcal{S}|} \alpha_t(i) = 1$ jointly with (13).

M Step. The values of the γ and ϵ computed during the E step can be used to maximise $Q(\Theta, \Theta^{old})$. Using (5), one obtains

$$q_i = \frac{\gamma_1(i)}{\sum_{i=1}^{|\mathcal{S}|} \gamma_1(i)} \tag{15}$$

and

$$a_{ij} = \frac{\sum_{t=2}^{T} \epsilon_t(i,j)}{\sum_{t=2}^{T} \sum_{j=1}^{|\mathcal{S}|} \epsilon_t(i,j)} \tag{16}$$

for the state prior and transition probabilities. The GMMs parameters become

$$\pi_{il} = \frac{\sum_{t=1}^{T} \gamma_t(i,l)}{\sum_{t=1}^{T} \gamma_t(i)}, \tag{17}$$

$$\mu_{il} = \frac{\sum_{t=1}^{T} \gamma_t(i,l) o_t}{\sum_{t=1}^{T} \gamma_t(i)} \tag{18}$$

and

$$\Sigma_{il} = \frac{\sum_{t=1}^{T} \gamma_t(i,l)(o_t - \mu_{il})^T (o_t - \mu_{il})}{\sum_{t=1}^{T} \gamma_t(i)} \tag{19}$$

where

$$\gamma_{il}(t) = \gamma_i(t) \frac{\pi_{il} b_{il}(o_t)}{b_i(o_t)}. \tag{20}$$

Eventually, the expert error probabilities are obtained using

$$p_i = \frac{\sum_{t|Y_t \neq i} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)} \tag{21}$$

and the annotations probabilities can be computed using (1).

The EM Algorithm. The EM algorithm can be implemented using the equations detailed above. Θ must be initialised before the first E step. This problem is already addressed in the literature for all the parameters, except d. A simple solution, used in this paper, consists in initialising d using

$$d_{ij} = \begin{cases} 1 - p_e & (i = j) \\ \frac{p_e}{|\mathcal{S}|-1} & (i \neq j) \end{cases} \tag{22}$$

where p_e is a small probability of expert annotation error. Equivalently, one can set $p_i = p_e$. For example, $p_e = .05$ is used in the experiments in Section 6.

5 ECG Segmentation

This section (i) quickly reviews ECG segmentation and the use of HMMs in this context and (ii) details the methodology used for the experiments in Section 6.

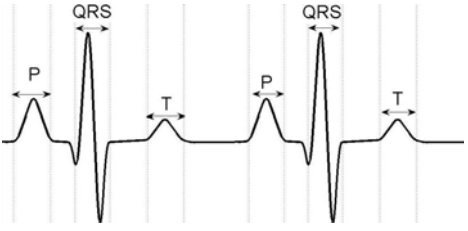

Fig. 3. Example of ECG signal, with annotations

5.1 ECG Signals

Electrocardiograms (ECGs) are periodic signals measuring the electrical activity of the heart. These time series are typically associated to a sequence of labels, called annotations (see Fig. 3). Indeed, physicians distinguish different kind of patterns called *waves*: P waves, QRS complexes and T waves. Moreover, physicians talk about baselines when the signal is flat, outside of waves. Here, only the B3 baseline between T and P waves is considered.

The ECG segmentation problem consists in predicting the labels for unlabeled observations, using the annotated part of the signal. Indeed, ECGs usually last for hours and it is of course impossible to annotate the entire signal manually.

In the context of ECG segmentation, (15) cannot be used directly. Indeed, only one ECG is available for HMM inference: the ECG of the patient under treatment. This is due to large inter-patient differences which prevent generalisation from one patient to the other. Here, q is simply estimated as the percentage of each observed annotation, as in the case of the supervised learning.

5.2 State of the Art

One of the most widely used, successful tool for ECG segmentation are HMMs [6, 15]. Typically, each ECG is firstly filtered using a 3-30 Hz band-pass filter. Then it is transformed using a continuous wavelet transform (WT) with an order 2 coiflet wavelet. The dyadic scales from 2^1 to 2^7 are kept in order to build the observations. Eventually, the resulting observations may be normalised component-wise, which is done in this paper. Fig. 4 shows the theoretical transitions in a HMM modelling an ECG.

Label noise has already been considered by [16] in the ECG context by using a semi-supervised approach. Annotations around boundaries are simply deleted,

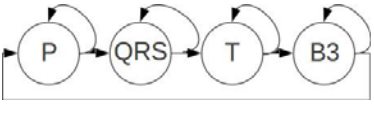

Fig. 4. Theoretical transitions in an ECG signal

which results in an intermediate situation between supervised learning and the Baum-Welch algorithm. Indeed, the remaining annotations are considered trustworthy and only the posteriors of the deleted annotations are estimated by EM. The width of the deletion window has to be selected, whereas this paper uses a noise model where the level of label noise is automatically estimated.

5.3 Experimental Settings

The two algorithms which are used for comparison are supervised learning and the Baum-Welch algorithm. Each emission model uses a GMM with 5 components. The EM algorithms are repeated 10 times and each repetition consists of at most 300 iterations. The initial mean of each GMM component is randomly chosen among the data in the corresponding class; the initial covariance matrix of each GMM component is set as a small multiple of the covariance matrix of the corresponding class.

Three classes of ECGs are used. Firstly, a set of 10 artificial ECGs are generated and annotated using the popular ECG waveform generator ECGSYN [4]. Secondly, 10 real ECGs are selected in the sinus MIT-QT database from Physiobank [5]. These Holter ECG recordings have been manually annotated by cardiologists with waveform boundaries for 30 to 50 selected beats in each recording. All recordings are sampled at 250 Hz. These ECGs were measured on real patients, but they are quite clean and easy to segment, for the patients were healthy. Thirdly, 10 ECGs are selected in the arrhythmia MIT-QT database from Physiobank [5]. These ECGs are more difficult to segment and the annotations are probably less reliable. Indeed, these patients were being treated for cardiac diseases and their ECG often differ significantly from the text-book ECGs. Only P waves, QRS complexes, T waves and B3 baselines are annotated.

Each ECG is segmented before and after the addition of artificial label noise. Different types and levels of artificial label noise are added to the annotations in order to test the robustness of each algorithm. The two label noises which are used here are called horizontal and uniform noise:

- The horizontal noise moves the boundaries of P and T waves by a random number of milliseconds drawn from a uniform distribution. This type of noise is particularly interesting in the context of ECG segmentation since it mimics the errors made by medical experts in practice. The uniform distribution used in the experiments is symmetric around zero and its half width is a given percentage of the considered wave duration.
- The uniform noise consist in randomly flipping a given percentage of the labels. This type of noise is the same as in previous experiments [1].

For each experiment, two measures are given: the average recall and precision. For each wave, the recall is the percentage of observations belonging to that wave which are correctly classified. The precision is the percentage of predicted labels which are correct, for a given label. Both measures are estimated using the ECGSYN annotations for the artificial ECGs, whereas human expert annotations are used for the real ECGs. In Section 5, recalls and precisions are systematically averaged over the four possible labels (P, QRS, T and B3).

For each algorithm, ECGs are split into training and test sets. The training set is used to learn the HMM, whereas the test set allows testing the HMM on independent data. For artificial ECGs, 10% of the signal is used for training, whereas the remaining 90% is used for test. For real ECGs, 50% of the signal is used for training, whereas the remaining 50% is used for test. This way, the size of the training sets are roughly equal for both artificial and real ECGs.

6 Experimental Results

This section compares the label noise-tolerant algorithm proposed in Section 4 to the two standard algorithms described in Section 3. The tests are carried out on three classes of ECGs, which are altered by two types of label noise. See Section 5 for more details about ECGs and the methodology used in this section.

6.1 Noise-Free Results

Tables 1 and 2 respectively show the recalls and precisions obtained on test beats for artificial, sinus and arrhythmia ECGs using supervised learning, Baum-Welch and the proposed algorithm. The annotations are the original annotations, without additional noise. Each ECG signal is segmented 40 times in order to evaluate the variability of the results. The results in the three first rows average the results of all runs for all ECGs, whereas other rows average the results of all runs for two selected ECGs. For the two selected ECGs, standard deviations are given. The standard deviations shown on the three first lines are the average of the standard deviations obtained for each ECG.

Results show that if one completely discards the available labels (i.e. with the Baum-Welch algorithm), the information loss is important. Indeed, the unsupervised algorithm always achieves significantly lower recalls and precisions. The results in terms of recall and precision are approximatively equal for the

Table 1. Recalls on original artificial, sinus and arrhythmia ECGs for supervised learning, Baum-Welch and the proposed algorithm

		supervised learning	Baum-Welch	proposed algorithm
average	artificial	95.21 ± 0.31	89.17 ± 2.20	95.14 ± 0.43
	sinus	95.35 ± 0.28	92.71 ± 1.60	95.60 ± 0.59
	arrhythmia	89.51 ± 0.69	82.48 ± 2.57	89.10 ± 0.90
ECG 1	artificial	94.98 ± 0.24	88.09 ± 2.33	94.87 ± 0.40
	sinus	95.75 ± 0.30	92.28 ± 1.50	96.44 ± 0.44
	arrhythmia	94.36 ± 0.46	81.77 ± 3.36	92.80 ± 1.79
ECG 2	artificial	93.34 ± 0.88	87.44 ± 3.17	93.50 ± 1.04
	sinus	95.88 ± 0.14	93.43 ± 0.74	96.01 ± 0.29
	arrhythmia	90.07 ± 0.35	88.01 ± 1.42	91.22 ± 0.46

Table 2. Precisions on original artificial, sinus and arrhythmia ECGs for supervised learning, Baum-Welch and the proposed algorithm

		supervised learning	Baum-Welch	proposed algorithm
average	artificial	95.53 ± 0.27	87.58 ± 3.06	95.34 ± 0.39
	sinus	95.86 ± 0.26	89.85 ± 2.23	94.38 ± 1.14
	arrhythmia	87.28 ± 0.75	77.37 ± 2.73	84.56 ± 1.45
ECG 1	artificial	95.51 ± 0.17	86.16 ± 3.10	95.14 ± 0.43
	sinus	96.81 ± 0.19	91.89 ± 1.50	95.96 ± 0.89
	arrhythmia	95.11 ± 0.96	72.13 ± 3.15	87.08 ± 4.03
ECG 2	artificial	94.76 ± 0.63	86.17 ± 4.26	94.67 ± 0.73
	sinus	96.42 ± 0.11	91.33 ± 1.87	95.82 ± 0.56
	arrhythmia	88.25 ± 0.32	86.14 ± 0.05	89.51 ± 0.49

proposed algorithm and supervised learning. One exception is the precision on arrhythmia signals, where the labels themselves are less reliable, making the performance assessment less reliable too.

6.2 Results with Horizontal Noise

Fig. 5, 6 and 7 show the results obtained for artificial, sinus and arrhythmia ECGs, respectively. The annotations are polluted by a horizontal noise, with the maximum boundary movement varying from 0% to 50% of the modified wave. For each figure, the first row shows the recall, whereas the second row shows the precision, both obtained on test beats. Each ECG signal is noised and segmented 40 times in order to evaluate the variability of the results. The curves in the first column average the results of all runs for all ECGs, whereas the curves in the second and third columns average the results of all runs for two selected ECGs. For the two last plots of each row, the error bars show the 95 % confidence interval around the mean on the 40 runs. The error bars shown on the first plot of each line are the average of the error bars obtained for each ECG.

Again, the performances of the unsupervised algorithm are the worst ones for small levels of noise. However, the results obtained using Baum-Welch seem to be less affected by the label noise. In most cases, the unsupervised algorithm achieves better results than supervised learning for large levels of noise. The effect of the noise level is probably due to the fact that the EM algorithm is initialised using the labelled observations. Therefore, the final result is also influenced by the label noise, for it depends on the initial starting point.

The performances of supervised learning and the label noise-tolerant algorithm are both affected by the increasing label noise. However, for large levels of noise, the label noise-tolerant algorithm achieves significantly better recalls and precisions than supervised learning. Supervised learning is only better in terms of precision for low levels of noise. Since the horizontal noise mimics errors made by medical experts, the above results suggest that using the proposed algorithm can improve the segmentation quality when the expert is not fully reliable.

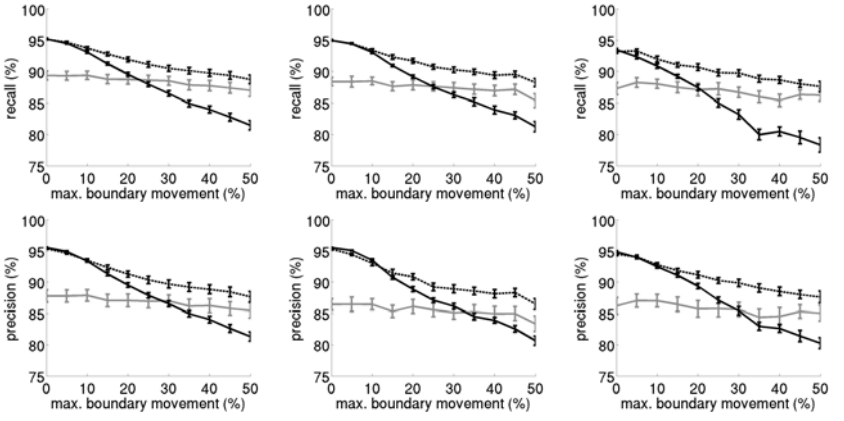

Fig. 5. Recalls and precisions on artificial ECGs with horizontal noise for supervised learning (black plain line), Baum-Welch (grey plain line) and the proposed algorithm (black dashed line), with respect to the percentage of the maximum boundary movement (0% to 50% of the modified wave). See text for details.

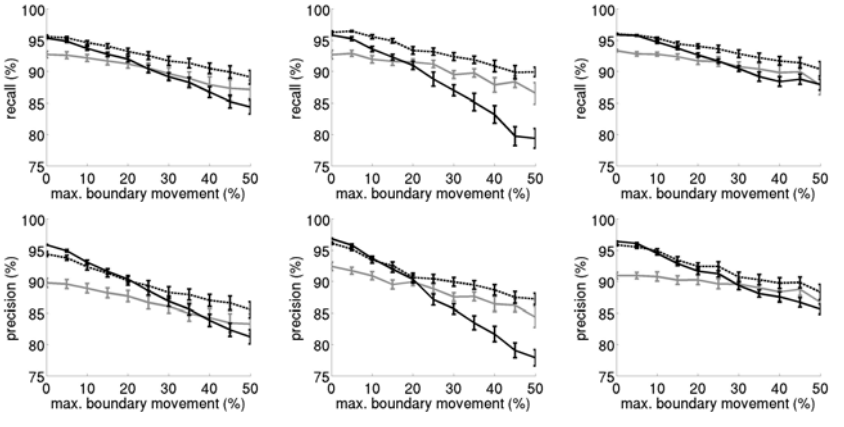

Fig. 6. Recalls and precisions on sinus ECGs with horizontal noise for supervised learning (black plain line), Baum-Welch (grey plain line) and the proposed algorithm (black dashed line), with respect to the percentage of the maximum boundary movement (0% to 50% of the modified wave). See text for details.

6.3 Results with Uniform Noise

Fig. 8, 9 and 10 shows the recalls and precisions obtained for artificial, sinus and arrhythmia ECGs, respectively. The annotations are polluted by a uniform noise, with a percentage of flipped labels varying from 0% to 20%. For each figure, the first row shows the recall, whereas the second row shows the precision, both obtained on test beats. Each ECG signal is noised and segmented 40 times in

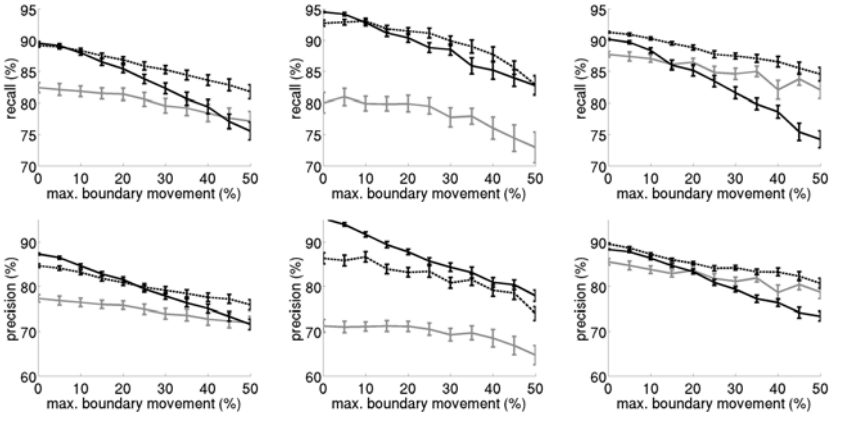

Fig. 7. Recalls and precisions on arrhythmia ECGs with horizontal noise for supervised learning (black plain line), Baum-Welch (grey plain line) and the proposed algorithm (black dashed line), with respect to the percentage of the maximum boundary movement (0% to 50% of the modified wave). See text for details.

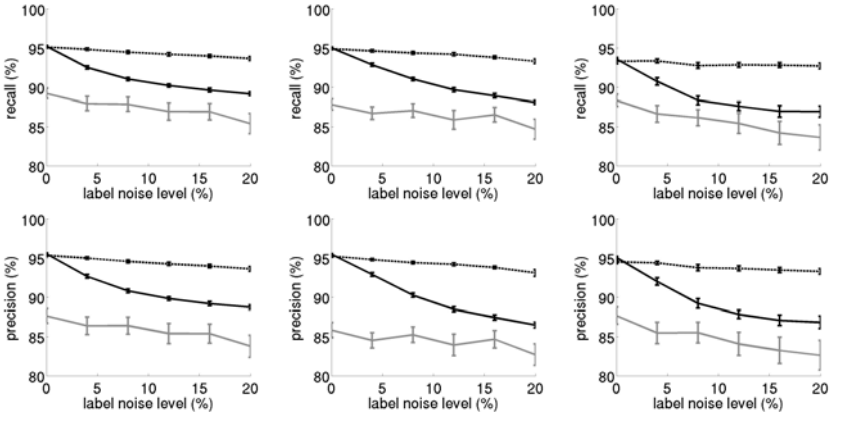

Fig. 8. Recalls and precisions on artificial ECGs with uniform noise for supervised learning (black plain line), Baum-Welch (grey plain line) and the proposed algorithm (black dashed line), with respect to the percentage of flipped labels (0% to 20%). See text for details.

order to evaluate the variability of the results. The curves in the first column average the results of all runs for all ECGs, whereas the curves in the second and third columns average the results of all runs for two selected ECGs. For the two last plots of each row, the error bars show the 95 % confidence interval around the mean on the 40 runs. The error bars shown on the first plot of each line are the average of the error bars obtained for each ECG.

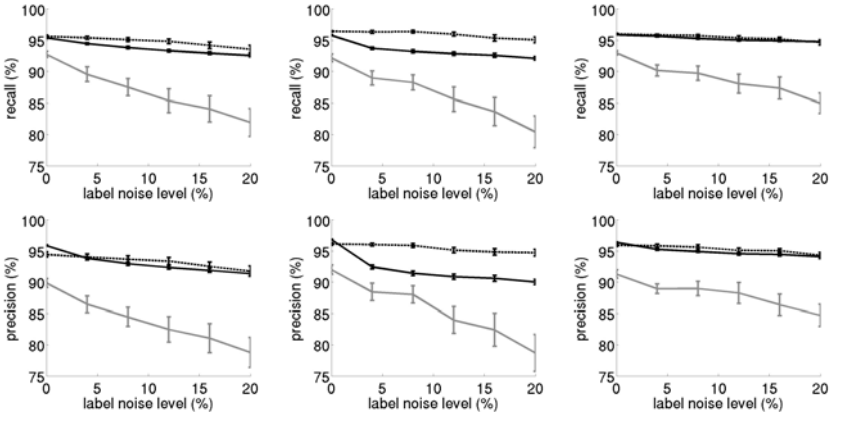

Fig. 9. Recalls and precisions on sinus ECGs with uniform noise for supervised learning (black plain line), Baum-Welch (grey plain line) and the proposed algorithm (black dashed line), with respect to the percentage of flipped labels (0% to 20%). See text for details.

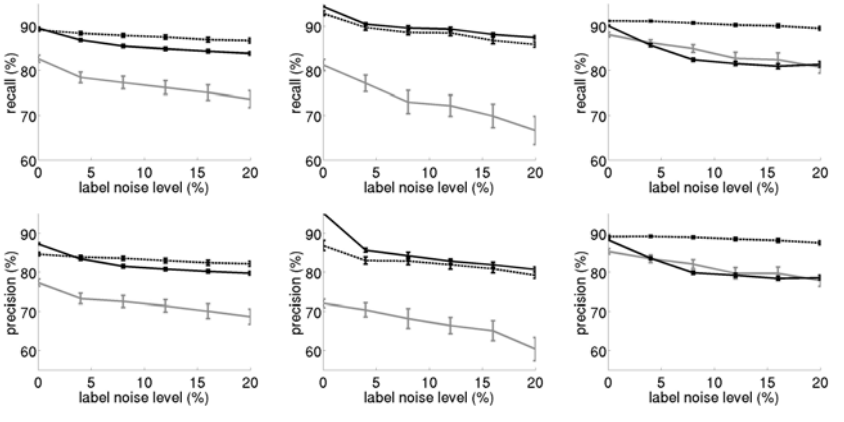

Fig. 10. Recalls and precisions on arrhythmia ECGs with uniform noise for supervised learning (black plain line), Baum-Welch (grey plain line) and the proposed algorithm (black dashed line), with respect to the percentage of flipped labels (0% to 20%). See text for details.

As for horizontal noise, the performances of Baum-Welch are significantly worse and decrease as the percentage of label noise increases. For the proposed algorithm, the recall and precision seem to be almost unaffected by the increasing level of label noise. For supervised learning, the recall and precision slowly decrease as the label noise increases. In terms of both recall and precision, the label noise-tolerant algorithm performs better than supervised learning when the level of noise is larger than 5%.

7 Conclusion

In this paper, a variant of the EM algorithm for label noise-tolerant HMM inference is proposed. More precisely, each observed label is assumed to be a noisy copy of the true, unknown state. The proposed EM algorithm relies on two steps to automatically estimate the level of noise in the set of available labels. First, during the E step, the posterior of the hidden state is estimated for each sample. Next, the M step computes the HMM parameters using the hidden true states, and not the noisy labels themselves, which results in a model which is less impacted by label noise.

Experiments are carried on both healthy and pathological ECGs signals artificially polluted by distinct types of label noise. Three types of inference algorithms for HMMs are compared: supervised learning, the Baum-Welch algorithm and the proposed noise-tolerant algorithm. The results show that the performances of the three approaches are adversely impacted by the level of label noise. However, the proposed noise-tolerant algorithm can yield better performances than the other two algorithms, which confirms the benefit of embedding the noise process into the inference algorithm. This improvement is particularly pronounced when the artificial label noise mimics errors made by medical experts, which suggests that the proposed algorithm could be useful when expert annotations are less reliable. The recall is improved for any label noise level, and the precision is improved for large levels of noise.

References

1. Lawrence, N.D., Schölkopf, B.: Estimating a kernel fisher discriminant in the presence of label noise. In: Proceedings of the Eighteenth International Conference on Machine Learning, ICML 2001, pp. 306–313. Morgan Kaufmann Publishers Inc, San Francisco (2001)
2. Li, Y., Wessels, L.F.A., de Ridder, D., Reinders, M.J.T.: Classification in the presence of class noise using a probabilistic kernel fisher method. Pattern Recognition 40, 3349–3357 (2007)
3. Bouveyron, C., Girard, S.: Robust supervised classification with mixture models: Learning from data with uncertain labels. Pattern Recognition 42, 2649–2658 (2009)
4. McSharry, P.E., Clifford, G.D., Tarassenko, L., Smith, L.A.: Dynamical model for generating synthetic electrocardiogram signals. IEEE Transactions on Biomedical Engineering 50(3), 289–294 (2003)
5. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 101(23), e215–e220 (2000)
6. Hughes, N.P., Tarassenko, L., Roberts, S.J.: Markov models for automated ECG interval analysis. In: NIPS 2004: Proceedings of the 16th Conference on Advances in Neural Information Processing Systems, pp. 611–618 (2004)
7. Brodley, C.E., Friedl, M.A.: Identifying Mislabeled Training Data. Journal of Artificial Intelligence Research 11, 131–167 (1999)

8. Barandela, R., Gasca, E.: Decontamination of training samples for supervised pattern recognition methods. In: Proceedings of the Joint IAPR International Workshops on Advances in Pattern Recognition, pp. 621–630. Springer, London (2000)
9. Guyon, I., Matic, N., Vapnik, V.: Discovering informative patterns and data cleaning. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) Advances in Knowledge Discovery and Data Mining, pp. 181–203 (1996)
10. Bootkrajang, J., Kaban, A.: Multi-class classification in the presence of labelling errors. In: Proceedings of the 19th European Conference on Artificial Neural Networks, pp. 345–350 (2011)
11. Côme, E., Oukhellou, L., Denoeux, T., Aknin, P.: Mixture model estimation with soft labels. In: Proceedings of the 4th International Conference on Soft Methods in Probability and Statistics, pp. 165–174 (2008)
12. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
13. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39(1), 1–38 (1977)
14. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics), 1st ed. 2006. corr. 2nd printing edition. Springer, Heidelberg (2007)
15. Clifford, G.D., Azuaje, F., McSharry, P.: Advanced Methods And Tools for ECG Data Analysis. Artech House, Inc., Norwood (2006)
16. Hughes, N.P., Roberts, S.J., Tarassenko, L.: Semi-supervised learning of probabilistic models for ecg segmentation. In: IEMBS 2004: Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 1, pp. 434–437 (2004)