Early Classification of Time Series by Hidden Markov Models with Set-Valued Parameters

Alessandro Antonucci
IDSIA, Manno, Switzerland
alessandro@idsia.ch

Mauro Scanagatta
IDSIA, Manno, Switzerland
mauro@idsia.ch

Denis Deratani Mauá
Universidade de São Paulo, São Paulo, Brazil
denis.maua@usp.br

Cassio Polpo de Campos
Queen's University Belfast, Belfast, UK
c.decampos@qub.ac.uk

Abstract

Hidden Markov models are popular tools for the modeling of multivariate time series. A set-valued quantification of the parameters might increase realism in the description of non-stationarity. A recent work shows that the computation of the bounds of the likelihood of a sequence with respect to such *imprecise* quantification can be performed in the same polynomial time of a precise computation with sharp values. This is the basis for a *credal* classifier of time series. For each training sequence we learn a set-valued model and compute the interval likelihood of the test sequence. The returned classes are those of the models associated to the undominated intervals. The approach is particularly accurate on the instances for which single classes are returned. We therefore apply this method to *early* classification of streaming data. As soon as the credal classifier returns a single output we assign a class label even if the stream is not terminated. Tests on a speech recognition benchmark suggest that the proposed approach might outperform a thresholding of the precise likelihoods with the classical models.

1 Introduction

Hidden Markov Models (HMMs) are popular probabilistic descriptions of paired sequences of states and observations [1], with applications in speech recognition [1] and computer vision [2], to name but a few. HMMs assume that the states have been generated by a first-order Markov Chain process, and that each observation has been generated based only on the paired state. The specification of an HMM comprises an *initial state probability distribution*, which specifies the probability that the process originates in a given state, a *state transition probability distribution*, which specifies the probability that the process will transit from a given state to another, and a *symbol emission probability distribution*, which specifies the probability of observing a symbol conditional on a state.

These parameters are often affected by severe uncertainty, as in the case of non-stationary data where no parametric form is known. A cautious approach is replacing the probability distributions with *credal sets* [3], i.e., closed and convex sets of probability distributions. Complete ignorance about a variable is represented as the credal set of all probability distributions. The *imprecise Dirichlet model* (IDM) learns credal sets from categorical data in a situation of near prior ignorance, providing a more reliable (although less informative) model of the underlying distribution than the more common Multinomial-Dirichlet model [4].

Efficient algorithms for inference with *imprecise hidden Markov models* (iHMMs), which allow the specification of a time- and state-discrete HMM with initial state, state transition and symbol emission credal sets in lieu of probability distributions have been recently developed in [5].

In particular, when the local parameters of an iHMM are specified as intervals, the bounds of the likelihood of a sequence can be computed with the same polynomial time complexity of a sharp likelihood computation in an HMM. In another paper [6], a procedure to learn iHMMs by combining the standard Baum-Welch algorithm with an IDM-based learning has been proposed.

Overall, this allows for iHMM-based time series classification. With HMMs a model is learned from each training sequence. The class label assigned to a test sequence is that of the model giving the highest likelihood to the sequence. Something similar can be done with iHMMs. The interval likelihoods associated to the training data are compared and the dominated intervals discarded. This implements a *credal* classifier possibly assigning more than a single class label to test sequences. Empirical validation in [5] displays the typical behaviour of credal classifiers, with single class labels assigned to “easy-to-classify” instances. In fact a standard HMM-based classifier is much more accurate on the instances for which the iHMM-based credal classifier returns a single output.

In this paper we exploit such a discriminative power of the credal classifier to perform *early* classification of time series [7]. The efficiency of the likelihood computation allows for an *online* analysis of the sequence (e.g., for footage data, the classifier is queried with the frame rate frequency). As soon as a single class label is returned we might reasonably expect that this is the actual class even if the sequence is not terminated yet. This is not so straightforward with HMMs whose sharp-valued likelihoods are such that there is always a dominating class.[1] To decide the class of the sequence before the end of the stream with HMMs, we need some threshold value to decide that a likelihood is clearly dominating the other ones. In the experiments we report here for a speech recognition benchmark the iHMM-based method is clearly outperforming HMMs, no matter which is the tuned choice of the threshold or criterion.

The paper is organized as follows. In Section 2 we give some detail about iHMMs with discussion about inference, learning, and extensions to support continuous data. In Section 3, we report the results of our experiments. Conclusions and outlooks are in Section 4.

2 Hidden Markov models with set-valued parameters

Precise HMMs A Hidden Markov model (HMM) describes a stochastic process over a sequence of *state variables* $Q_1, \ldots, Q_T$ and *manifest variables* $O_1, \ldots, O_T$. Each state variable Q_t, $t = 1, \ldots, T$, takes values in a finite set $\mathcal{Q} = \{1, \ldots, N\}$; each manifest variable O_t takes value in a finite set $\mathcal{O} = \{1, \ldots, M\}$. We initially consider the case of discrete manifest variables. The extension to the continuous case will be discussed later. We denote an arbitrary value of state variable Q_t by q_t, i or j, and similarly for O_t. The parameter t that indexes either family of variables is called *time*. The stochastic process satisfies two properties: (P1) A state variable Q_t is independent of all the variables in the past given its immediate predecessor state variable Q_{t-1}, that is, $\Pr(Q_t = q_t | Q_{1:t-1} = q_{1:t-1}, O_{1:t-1} = o_{1:t-1}) = \Pr(Q_t = q_t | Q_{t-1} = q_{t-1})$, where the notation $X_{1:r} = x_{1:r}$ denotes the event $X_1 = x_1, \ldots, X_r = x_r$. (P2) A manifest variable O_t is stochastically independent of any other variable given the state variable Q_t: $\Pr(O_t = o_t | O_{1:t-1} = o_{1:t-1}, O_{t+1:T} = o_{t+1:T}, Q_{1:T} = q_{1:T}) = \Pr(O_t = o_t | Q_t = q_t)$. A hidden Markov model can be therefore regarded as a tuple $\lambda = (a_2^1, \ldots, a_T^N, b_1^1, \ldots, b_T^N, \pi)$ where $a_t^i(j) := \Pr(Q_t = j | Q_{t-1} = i)$ with $i = 1, \ldots, N$ and $t = 2, \ldots, T$, $b_t^i(j) := \Pr(O_t = j | Q_t = i)$ with $i = 1, \ldots, N$ and $t = 1, \ldots, T$, and $\pi(i) := \Pr(Q_1 = i)$ with $i = 1, \ldots, N$. The functions a_t^i, b_t^i and π are called the *transition, emission and initial probability distributions*, respectively. The model is said to be *stationary* if for any i, t and t' we have that $a_t^i = a_{t'}^i$ and $b_t^i = b_{t'}^i$. An HMM λ is a succinct representation of a stochastic process satisfying Properties P1 and P2 defining the following joint probability distribution

$$p_\lambda(q_{1:T}, o_{1:T}) := \Pr_\lambda(Q_{1:T} = q_{1:t}, O_{1:T} = o_{1:T}) = \pi(q_1) b_1^{q_1}(o_1) \prod_{t=2}^{T} a_t^{q_{t-1}}(q_t) b_t^{q_t}(o_t) \,. \quad (1)$$

Given a sequence $o_{1:T}$, the computation of its likelihood according to λ, i.e.,

$$p_\lambda(o_{1:T}) := \sum_{q_{1:T}} p_\lambda(q_{1:T}, o_{1:T}) \quad (2)$$

can be efficiently performed in time $O(TN^2)$ by dynamic programming [1].

[1] Apart from the very unlikely case of two or more HMMs assigning the same likelihood to a sequence.

Credal sets In HMMs uncertainty about the initial state, state transition, and symbol emission are modeled by single probability measures. This requirement might lead to an inaccurate description. *Credal* sets, i.e., convex and closed sets of probability measures can provide a more adequate representation of knowledge. We denote by $\mathbb{K}_{X_{1:r}}$ a credal set of probability measures over $X_{1:r}$. The set ext$\mathbb{K}_{X_{1:r}}$ denotes the extreme functions of the $\mathbb{K}_{X_{1:r}}$. A credal set can be specified by linear inequalities over the probabilities of the singletons. A simple and commonly used type of inequalities specifying a credal set over X is of the form $0 \leq \ell(x) \leq p(x) \leq u(x) \leq 1, \quad \forall x \in \mathcal{X}$, where ℓ and u are functions from $\mathcal{X}$ to $[0,1]$. Credal sets characterized in such a way are said to be specified by interval-valued probabilities. Given a credal set $\mathbb{K}_{X_{1:r}}$, X_i and X_j are *strongly independent* given $X_k = x_k$ if they are independent under every probability measure in ext$\mathbb{K}_{X_{1:r}}$ [8].

Imprecise hidden Markov models An imprecise hidden Markov model (iHMM) is a concise description of the same stochastic process described by an HMM, except that we replace the representation of uncertainty using single probability measures by credal sets and the notion of stochastic independence by strong independence. An imprecise hidden Markov model (iHMM) is therefore a tuple $\Lambda = (A_2^1, \ldots, A_T^N, B_1^1, \ldots, B_T^N, \Pi)$ where $A_t^i := \mathbb{K}_{Q_t}^{Q_{t-1}=i}$, $B_t^i := \mathbb{K}_{O_t}^{Q_t=i}$ and $\Pi := \mathbb{K}_{Q_1}$ with the same ranges for the indexes as in the HMM definition. The sets A_t^i, B_t^i and Π are called *transition, emission and initial sets*. An iHMM is said to be *homogeneous* if for all i: (i) transition credal sets $A_2^i, \ldots, A_T^i$ are equal, and (ii) emission credal sets $B_1^i, \ldots, B_T^i$ are equal. An iHMM Λ induces a credal set $\mathbb{K}_{Q_{1:T}, O_{1:T}}$ over $(Q_{1:T}, O_{1:T})$ whose *extreme distributions* obey the factorization in Eq. (1) for each $a_2^1 \in \text{ext}A_2^1, \ldots, a_T^N \in \text{ext}A_T^N, b_1^1 \in \text{ext}B_1^1, \ldots, b_T^N \in \text{ext}B_T^N, \pi \in \text{ext}\Pi$. Given an iHMM Λ and a sequence $o_{1:T}$, the HMM task in Eq. (2) generalises to iHMMs as follows:

$$\underline{p}_\Lambda(o_{1:T}) := \min_{\lambda \in \Lambda} p_\lambda(o_{1:T}) = \min_{\pi \in \Pi, a_t^i \in A_t^i, b_t^j \in B_t^j} \sum_{q_{1:T}} \pi(q_1) b_1^{q_1}(o_1) \prod_{t=2}^{T} a_t^{q_{t-1}}(q_t) b_t^{q_t}(o_t)\,, \tag{3}$$

and analogously for the upper bound $\overline{p}_\Lambda$. We call these bounds upper and lower likelihoods. In [5], it is shown that when credal sets are specified by interval-valued probabilities the total running time is $O(TN^2)$ as in the precise case in Eq. (2).[2]

Coping with continuous variables Many HMM applications have continuous manifest variables, i.e., $\mathcal{O} := \mathbb{R}$. Normality is therefore assumed in the emission terms, i.e., $b_t^{q_t}(o_t) := \mathcal{N}_{\mu(q_t)}^{\sigma(q_t)}(o_t)$.[3] The computation in Eq. (2) and, with precise emission terms, Eq. (3) can be extended to the continuous case by discretization. Given the observed sequence $o_{1:T}$, an interval $\mathcal{I}_i := [o_i - \epsilon, o_i + \epsilon] \subset \mathbb{R}$ is defined for each $i = 1, \ldots, T$. For small ϵ, these intervals do not overlap, unless there are identical observations in $o_{1:T}$ (intervals coincide in that case). A partition to define the discretization $\tilde{O}_t$ of O_t is therefore obtained. Theorems in [9] for Bayesian (HMMs) and credal (iHMMs) networks, allows to equivalently regard the observed variable $\tilde{O}_t$ as a Boolean variable $\tilde{O}_t'$ whose true state is the actual one, i.e., $\{o_i \in \mathcal{I}_i\}$. The emission term becomes therefore $b_t^{q_t}(\tilde{O}_t' = 1) := \int_{o_t - \epsilon}^{o_t + \epsilon} b_t^{q_t}(o_t)\mathrm{d}t$. This holds for any finite ϵ, and remains true for $\epsilon \to 0$, as the likelihood is a linear function of the emission terms. Overall, the likelihood (density) of a sequence corresponds to that of a discrete sequence in a non-stationary HMM (or iHMM) with Boolean manifest variables and $b_t^{q_t}(\tilde{O}_t' = 1) := \mathcal{N}_{\mu(q_t)}^{\sigma(q_t)}(o_i)$.[4]

Learning iHMMs In HMMs incomplete data are processed with the Baum-Welch algorithm [1], which implements an EM procedure to infer the parameters in the absence of observations of the state variables. The Imprecise Dirichlet Model (IDM) [4] is a technique to learn interval-valued credal sets from *complete data*. Fractional Baum-Welch counts are regarded as the result of pseudo-observations of the state variables and used as an input for the IDM [6]. For state transitions this corresponds to:

$$\frac{E[n(Q_{t-1}\!=\!i, Q_t\!=\!j)]}{\sum_j E[n(Q_{t-1}\!=\!i, Q_t\!=\!j)] + s} \leq a_t^i(j) \leq \frac{E[n(Q_{t-1}\!=\!i, Q_t\!=\!j)] + s}{\sum_j E[n(Q_{t-1}\!=\!i, Q_t\!=\!j)] + s}, \tag{4}$$

where $E[n(Q_{t-1} = i, Q_t = j)]$ are the expected counts for a transition from state i to state j obtained through Baum-Welch algorithm and s is the equivalent sample size of the Dirichlet priors.

[2] An implementation of this algorithm is available at `https://github.com/denismaua/ihmm`.

[3] $\mathcal{N}_\mu^\sigma(x)$ denotes a Gaussian over x with mean μ and variance σ^2.

[4] A rescaling is applied for density values bigger than one.

3 Early credal classification of time series

Credal classifiers HMMs are used to classify sequential data by choosing the model that best fits a sequence according to the likelihood. E.g., in speech recognition, a different HMM is learned for each individual; the speaker (i.e., the class) is determined by selecting the HMM maximizing the probability of a recorded passphrase (i.e., a sequence). If $\{\lambda_j\}_{j=1}^K$ are the HMMs associated to the training set, a test sequence $o_{1:T}$ is labeled with the class of $\lambda^* = \mathrm{argmax}_{j=1,\ldots,K}\, p_{\lambda_j}(o_{1:T})$.

Given two iHMMs Λ_1 and Λ_2 and a sequence $o_{1:T}$, Λ_1 *dominates* Λ_2 for $o_{1:T}$, denoted $\Lambda_1 \succ \Lambda_2$, if and only if $\underline{p}_{\Lambda_1}(o_{1:T}) > \overline{p}_{\Lambda_2}(o_{1:T})$. Dominance suggests that iHMMs can be used as *credal classifiers* [10] for reliable/robust sequence classification in the same way as HMMs are used for classifying sequential data. A class label associated to model Λ_1 is preferred as a classification of $o_{1:T}$ over a class label associated to model Λ_2 if and only if $\Lambda_1 \succ \Lambda_2$. Given a finite set $\Lambda_1, \ldots, \Lambda_K$ of iHMMs, credal classification outputs the set of undominated models $\Lambda^* = \{\Lambda_k : \nexists j \text{ such that } \Lambda_j \succ \Lambda_k\}$.

Early detection Both the standard (HMM-based) and the credal (iHMM-based) classifier can efficiently process streaming test data by evaluating the sequence at any time step. An early detection with iHMMs is performed if the credal classifier returns a single output before the end of the stream. HMMs always return a single output. An early detection with these model is performed only if the ratio between the highest and the second-highest likelihood exceeds a threshold $\eta > 1$ [11]. To compare these methods we add to a zero/one loss function a term $\rho\tau$, linear in the number τ of time steps elapsed before the recognition (ρ is the loss of a single time step elapsed). Condition $\rho T \ll 1$, where T is the sequence length, makes late-but-right detections better than wrong-but-prompt ones.

Experiments We evaluate the two approaches on the *Japanese Vowels* speaker recognition dataset [12]. 270 sequences representing sound records from nine male speakers are described by 12 continuous features. To test early recognition we only consider the first seven time steps, this being the maximum length of a training sequence. To avoid discretization, precise emission terms are specified. IDM constraints as in Eq. (4) are used to define the transition credal sets, while Π is vacuous. In the learning of the models we set $s = 2$ and $N = 2$. Nine sequences, one for each speaker uttering the same sound, are used for the training and the rest for testing. A 30-fold cross validation scheme is used. Fig. 1 compares the HMM and iHMM losses for different values of the loss parameter ρ and the threshold η. The results are clear, the difference is always positive, this meaning that iHMMs outperform HMM on all the considered configurations. The difference increases for increasing values of η (making the HMM more cautious), and ρ (making waiting more time steps less expensive).

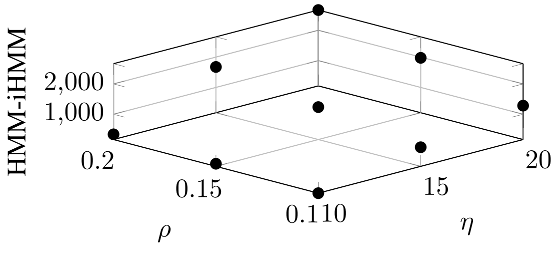

Figure 1: Difference between HMM and iHMM losses in the Japanese Vowels dataset.

4 Conclusions and outlooks

A novel approach to early classification of time series by HMMs with set-valued parameters is considered. Preliminary results are promising: the set-valued approach outperforms the standard one on a speech recognition benchmark. Besides an empirical validation against other methods (e.g., [13]), we intend to evaluate the effects of an imprecise quantification in the emission terms by using the models in [14].

References

[1] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, IEEE 77 (2) (1989) 257–286.

[2] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, Transactions on Pattern Analysis and Machine Intelligence 29 (12) (2007) 2247–2253.

[3] I. Levi, The Enterprise of Knowledge, MIT Press, 1980.

[4] P. Walley, Inferences from multinomial data: Learning about a bag of marbles, Journal of the Royal Statistical Society. Series B (Methodological) 58 (1) (1996) 3–57.

[5] D. Mauá, C. de Campos, A. Antonucci, Hidden Markov models with set-valued parameters, Neurocomputing (in press). doi:10.1016/j.neucom.2015.08.095.

[6] A. Antonucci, R. De Rosa, A. Giusti, F. Cuzzolin, Robust classification of multivariate time series by imprecise hidden Markov models, International Journal of Approximate Reasoning 56 (B) (2015) 249–263.

[7] Z. Xing, J. Pei, P. Yu, Early prediction on time series: A nearest neighbor approach, in: Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-09), 2009.

[8] F. Cozman, Credal networks, Artificial Intelligence 120 (2000) 199–233.

[9] A. Antonucci, A. Piatti, Modeling unreliable observations in Bayesian networks by credal networks, in: L. Godo, A. Pugliese (Eds.), Proceedings of SUM 2009, Vol. 5785 of Lecture Notes in Computer Science, Springer, 2009, pp. 28–39. doi:10.1007/978-3-642-04388-8_4.

[10] M. Zaffalon, E. Fagiuoli, Tree-based credal networks for classification, Reliable Computing 9 (6) (2003) 487–509.

[11] N. Hatami, C. Chira, Classifiers with a reject option for early time-series classification, in: Proceedings of the IEEE Symposium on Computational Intelligence and Ensemble Learning, CIEL 2013, IEEE Symposium Series on Computational Intelligence (SSCI), 16-19 April 2013, Singapore, 2013, pp. 9–16.

[12] M. Kudo, J. Toyama, M. Shimbo, Multidimensional curve classification using passing-through regions, Pattern Recognition Letters 20 (1999) 1103–1111.

[13] M. F. Ghalwash, Z. Obradovic, Early classification of multivariate temporal observations by extraction of interpretable shapelets, BMC Bioinformatics 13 (2012) 195.

[14] A. Benavoli, M. Zaffalon, Prior near ignorance for inferences in the k-parameter exponential family, Statistics doi:10.1080/02331888.2014.960869.