Time Series Classification by Imprecise Hidden Markov Models

Alessandro ANTONUCCI [a,1] and Rocco DE ROSA [b],
[a] *IDSIA, Galleria 2, Via Cantonale, CH-6928 Manno-Lugano (Switzerland)*
alessandro@idsia.ch
[b] *Computer Science Dept., University of Milan, Via Comelico 39/41, 20135 Milan, Italy*
rocco.derosa@studenti.unimi.it

Abstract. Hidden Markov models (HMMs) are powerful tools for modelling the generative and observational processes behind time series. For short sequences, the small amount of data can make unreliable the estimates returned by the EM algorithm, which is generally used to learn HMMs. To gain robustness in these cases, an *imprecise* version of the EM algorithm, achieving an interval-valued quantification of the model parameters can be considered instead. The bounds of the likelihood assigned to a particular sequence with respect to these intervals can be efficiently computed. Overall, this provides a time series classification algorithm. To classify a new sequence, the bounds of the likelihood associated to the HMMs learned from the supervised sequences are evaluated, and the returned class label is that of the highest-likelihood interval. If two or more of these intervals overlap and they are associated to different labels, the classifier returns multiple classes, this corresponding to a condition of partial indecision for the class of a particular sequence. An application to human action recognition shows the effectiveness of this approach in discriminating the hard-to-classify instances (those for which the classifier returns many classes) from the "easy" ones (those for which a single class, which mostly is the correct one, is returned). This suggests the opportunity of an application of the proposed approach as an useful preprocessing tool for other time series classifiers.

Keywords. Time Series, Hidden Markov Models, Credal Networks, EM Algorithm, Imprecise Dirichlet Model, Imprecise Probability.

Introduction

Supervised classification of temporally structured data is an important branch of machine learning with lots of applications in many different fields, ranging from recognition problems (*e.g.*, actions, gestures, speech) to finance (*e.g.*, changes in stock markets), robotics (*e.g.*, detecting temporal events from sensors) and medicine (*e.g.*, cardiology), just to mention a few. A straightforward application of standard classification algorithms to process data of this kind is clearly possible, but it typically leads to poor performances.

For a better modelling, and hence higher classification accuracies, it seems to be crucial to take into account the temporal correlation among these data, and then adopt

[1]Corresponding Author: Alessandro Antonucci, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, Galleria 2 - Via Cantonale, CH-6928 Manno-Lugano (Switzerland); e-mail: alessandro@idsia.ch

a dynamic model. In the probabilistic framework, this is generally done by learning a Hidden Markov Model (HMM) from every time series by the *Expectation Maximisation* (EM) algorithm [8]. Yet, the estimates returned by the EM algorithm for the HMM parameters when the series are short, and hence only few data are available, are known to be potentially unreliable. This suggests the opportunity of a more robust learning procedure to be adopted in these cases. An interesting direction to tackle this issue is represented by the theory of *imprecise probability* [9], where uncertainty is described by convex sets (instead of single) probability distributions. In particular, when learning probabilistic models from multinomial data, the so-called *imprecise Dirichlet model* (IDM, [10]) allows for learning an imprecise-probabilistic model by a Bayesian-like approach, where, instead of a single, a set of priors modelling a condition of near-ignorance about the model parameters is adopted.

Here, we apply these ideas to the quantification of a HMM by means of the EM algorithm. In our approach, this is simply achieved by considering the expectations about the configurations of the hidden variables of the HMM, as computed by the forward/backward algorithms after each iteration of the algorithm. These expectations are regarded as (non-integer) counts for the corresponding variables, with IDM used to learn from them an imprecise, interval-valued, quantification of the transition probabilities for the hidden variables of the HMM. This provides a generally more reliable and robust approach to the quantification of the HMM parameters.

As well as standard HMMs can be regarded as particular (dynamic) Bayesian networks [7], the interval-valued quantification of the HMM achieved by the imprecise EM transforms the model into a *credal network* [3]. Inference in credal networks is typically more demanding than in the Bayesian case [4]; yet the evaluation of the probability assigned by an imprecise HMM to an observable sequence (*i.e.*, the *likelihood* of the observable data) can be efficiently performed by the message propagation algorithm proposed in [11]. Of course, as the model is quantified by sets of distributions (those consistent with the interval constraints), only the lower and upper bounds of the likelihood can be evaluated. These descriptors can be used to perform classification. The likelihood intervals computed for the test sequence over the different HMMs learned from the supervised sequences are compared. In particular the highest-likelihood intervals are considered: if they all are associated to the same class label, the classifier returns a single class, while multiple classes are returned when intervals associated to different labels overlap. The latter case corresponds to a condition of (partial or complete) indecision about the label to assign to the time series under consideration.

The performances of this classifier are empirically tested in a computer vision problem involving the recognition of human actions. The experiments show the ability of our approach to separate the hard-to-classify instances from the easy ones. In practice, when the classifier returns a single action label, this is very likely to be the correct one. On the other side, when more than a single class is returned, the sequence could be particularly hard to be identified. Yet, the correct class label is generally included into the set of options returned by the classifier.

The paper is organised as follows: in Sect. 1 we briefly review the necessary background material about HMMs, EM and imprecise probabilities; then, in Sect. 2, we propose our imprecise-probabilistic version of the EM algorithm for HMMs; this approach is applied to time series classification in Sect. 3 and tested on a human action recognition task in Sect. 4; conclusions and outlooks are finally discussed in Sect. 5.

1. Background

1.1. Hidden Markov Models (HMMs)

HMMs [8] are very popular dynamic probabilistic models intended to describe temporal sequences when coping with uncertainty. The *hidden* layer of the model is a collection of (directly unobservable) variables, one for each (discrete) time step, modelling the actual state of the system. This is *Markovian*, which means that each configuration is only affected by that at the previous time step. The *observable* layer corresponds to a second collection of variables, again one for each time step. These are intended to report observable information about the system configuration. The configuration of an observable variable is only affected by the relative hidden variable.

If time t varies in $\{1, \ldots, T\}$ (*i.e.*, we have T discrete time steps), the model is defined over variables $\{(X_t, \mathbf{O}_t)\}_{t=1}^{T}$, where the hidden variable X_t takes values in a finite set $\mathcal{X}$, which is the same at any time ($|\mathcal{X}| = N$); F real-valued features are observed, *i.e.*, the observable variables $\mathbf{O}_t$ are assumed to take values in $\mathbb{R}^F$. We denote by O_t^f the f-th feature (*i.e.*, coordinate of $\mathbf{O}_t$) of the model. The independence assumptions characterising the model induce in the joint density $P(x_1, \ldots, x_T, \mathbf{o}_1, \ldots, \mathbf{o}_T)$ the following factorisation:

$$P(x_1) \prod_{t=1}^{T-1} P(x_{t+1}|x_t) \prod_{t'=1}^{T} \prod_{f=1}^{F} \mathcal{N}(o_{t'}^f)_{\mu^f(x_{t'}),\sigma^f(x_{t'})}, \tag{1}$$

where $\mathcal{N}(o)_{\mu,\sigma}$ is a Gaussian distribution over o with mean μ and standard deviation σ, $\mu^f(x)$ is the mean of the Gaussian distribution associated to the f-th feature, when the corresponding hidden variable is in the state $x \in \mathcal{X}$, and similarly for the standard deviation. Means, standard deviations, and the hidden-to-hidden transitions are independent of t and the features are real-valued: thus, (1) defines a continuous stationary *hidden Markov model*. We denote by λ a generic HMM specification.

1.2. Learning HMMs by EM Algorithm

When complete data about both hidden variables and observable features are available, maximum-likelihood estimators can be used for HMM quantification. Yet, hidden variables are by definition unobservable, and the available temporal sequence corresponds to the features observations only, *i.e.*, $\{\mathbf{O}_t\}_{t=1}^{T}$. Algorithms to learn the model from these incomplete data should be therefore considered instead. A typical choice is the *expectation maximisation* (EM, [6]): a recursive estimation, which, after a random initialisation, converges to a local maximum of the likelihood.

For HMMs, closed formulae can be obtained for each recursion. *E.g.*, for the probabilities on the hidden layer:

$$P^{(\mathrm{new})}(x_1) = \frac{P^{(\mathrm{old})}(x_1, \mathbf{o}_1, \ldots, \mathbf{o}_T)}{\sum_{x_1 \in \mathcal{X}} P^{(\mathrm{old})}(x_1, \mathbf{o}_1, \ldots, \mathbf{o}_T)}, \tag{2}$$

$$P^{(\mathrm{new})}(x_{t+1}|x_t) = \frac{\sum_{t=1}^{T-1} P^{(\mathrm{old})}(x_t, x_{t+1}, \mathbf{o}_1, \ldots, \mathbf{o}_T)}{\sum_{t=1}^{T-1} P^{(\mathrm{old})}(x_t, \mathbf{o}_1, \ldots, \mathbf{o}_T)}, \tag{3}$$

where the right-hand sides of these equations are efficiently computed by the forward-backward algorithm [8], and they can be regarded as ratios of expected number of hidden variables in a given state or doing a particular transition. Yet, these estimates are sometimes unreliable and unstable with respect to the initialisation of the parameters, especially when only few data are available.[2]

1.3. Imprecise Probability and Imprecise HMMs

The theory of *imprecise probability* [9] is a generalisation of the classical Bayesian theory of probability, where instead of single probability distributions, sets of distributions are assumed to provide a more robust and realistic model of uncertainty. In particular, uncertainty about the state of a variable X is described by a *credal set* $K(X)$, which is a collection of distributions $P(X)$ over X. Inference over a credal set is intended as the computation of the lower and upper bounds, with respect to the whole set of distributions, of the considered expectation. This problem can be solved by considering only the *extreme* points of the credal set, this transforming an optimisation over a continuous domain into a combinatorial task.

The *imprecise Dirichlet model* (IDM, [10]) can be used to learn a credal set $K(X)$ from data. This set is made of all the distributions $P(X)$ consistent with the constraints:

$$\frac{n(X=x)}{N+s} \leq P(X=x) \leq \frac{n(X=x)+s}{N+s}, \tag{4}$$

for each $x \in \mathcal{X}$, where $n(X=x)$ is the number of records such that $X=x$, N the total amount of data, and the hyperparameter s describes the degree of caution in the inferences.

Credal sets have been adopted to extend Bayesian nets to imprecise probabilities. The result is a class of imprecise probabilistic graphical models called *credal nets* [3]. Thus, as well as a standard HMM can be regarded as a Bayesian net, a HMM where the "precise" prior $P(X_1)$ and transition matrix $P(X_{t+1}|X_t)$ have been replaced by a corresponding collection of credal sets, is an *imprecise HMM*, which corresponds to a particular credal net. Despite the hardness characterising inference in general credal nets [4], a number of algorithms to efficiently compute exact inferences has been developed for same special cases. For our purposes, it is worth noting that the algorithm in [11] can efficiently compute the bounds of the marginal probability for a set of variables in a tree-shaped credal net.

2. An Imprecise-Probabilistic EM

As noted in Sect. 1.2, EM estimates when only few data are available can be inaccurate. Thus, on the basis of the discussion in Sect. 1.3, in these cases, it seems quite natural to gain robustness by formulating an imprecise-probabilistic version of the EM algorithm.

[2]The missing data make the likelihood function not concave and not unimodal. Thus, generally speaking, there could be multiple maximum-likelihood estimators, and even if the EM would converge to the global maximum, the estimated parameters could be completely different from those generating the data. It seems reasonable to conjecture that this issue is particularly important when only few (incomplete) data are available.

This becomes particularly simple by exploiting the fact that recursions in (2) and (3) have been already formalised in terms of expected counts. Thus, we obtain interval-valued estimates by replacing the integer counts required by the IDM as in (4) with the corresponding expected counts. This provides a sufficient level of cautiousness, which is clearly advisable when coping with few data. We call this approach *imprecise EM*.[3]

On the basis of the above presented idea, we can easily achieve an imprecise quantification for an HMM, or, in other words, we can learn an imprecise HMM from any observable sequence. Note that the imprecise quantification regards only the probabilities for the hidden variables, while for the observable variables we keep the precise quantification returned by the standard EM.[4] In order to see how this technique in practice, let us consider the following simple example.

Example 1 Consider a HMM defined as in Sect. 1.1, with $N = 2$, $P(X_1 = 1) = .5$, $P(X_{t+1} = 1|X_t = 0) = .7$, $P(X_{t+1} = 1|X_t = 1) = .5$, and a single feature $\sigma = .1$ and $\mu(X = 0) = -1$, $\mu(X = 1) = 1$. This HMM is used to generate an observable sequence $\{o_t\}_{t=1}^{T}$. Both the standard EM and the imprecise version proposed here are used to learn the parameter $P(X_{t+1} = 1|X_t = 0)$. The results for different values of T are in Figure 2.

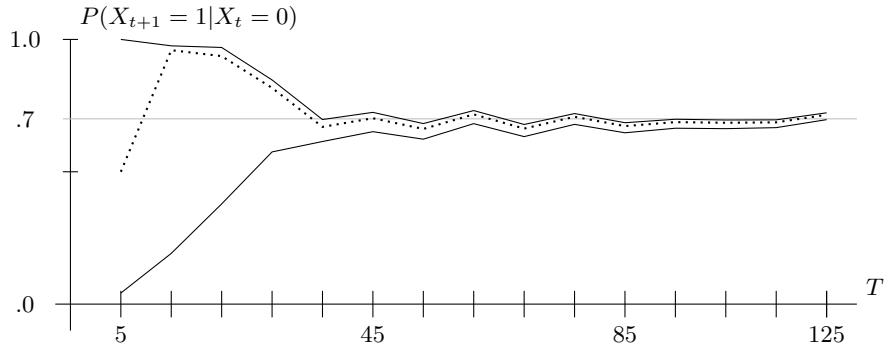

Figure 1. Imprecise vs. precise EM. The lower and upper bounds returned by the imprecise EM (continuous lines) are compared with the precise estimate provided by the standard EM (dotted line) and with the true numerical value of the model used to generate the data (grey line).

According to this example, the imprecise EM seems to manifest the behaviour we expect: for large amount of data it basically collapses into the standard EM estimates, while with fewer data a more robust estimate is provided, this corresponding to a probability intervals which generally includes the true value of the parameter (and, of course, the precise estimate). Considering the structure of the IDM as in (4) and the fact that for longer chains the expectations for the counts should increase, this should happen also in the general case.

[3] Clearly, this is just a possible, simple, approach to the generalisation towards the imprecise-probabilistic framework of the EM algorithm. A more sophisticated approach would require a Bayesian formulation algorithm, and then its imprecise extension by considering a set of priors as in the IDM.

[4] Learning imprecise-probabilistic models for continuous variables from data is a problem less studied in the literature. Despite some recent advances in this field, for the moment we prefer to put imprecision only on the hidden layer of the HMM.

3. Supervised Time Series Classification by Imprecise HMMs

Given a collection of time series, the EM algorithm can be used to learn a HMM for each sequence, and, analogously, the imprecise version of the EM proposed in Sect. 2, can be used to learn an imprecise HMM. Let $\mathcal{C}$ denote a set of (mutually exclusive and exhaustive) class labels, which can be attached to the time series under consideration. Given a labeled collection of time series $\{(c_d, s_d)\}_{d=1}^{D}$, let us denote respectively by λ_d and $\overline{\underline{\lambda}}_d$ the precise and imprecise HMMs obtained from sequence s_d, and $c_d \in \mathcal{C}$ the relative class label. To recognise a new unlabeled sequence s_{D+1} in the precise framework, we can simply identify the class label c_{d^*} of the HMM assigning highest likelihood to the sequence under consideration, *i.e.*,

$$d^* := \arg\max_{d=1,\ldots,D} P(\mathbf{o}_1^{D+1}, \ldots, \mathbf{o}_T^{D+1} | \lambda_d). \tag{5}$$

To extend (5) to imprecise probabilities, we first note that, for the likelihood assigned to an observable sequence by the imprecise HMM, we can only evaluate lower and upper bounds. This corresponds to an inference (marginalisation) problem on a credal net with tree topology. The problem can be efficiently solved by the algorithm in [11], this making possible to obtain $\underline{P}(\mathbf{o}_1^{D+1}, \ldots, \mathbf{o}_T^{D+1} | \overline{\underline{\lambda}}_d)$, and similarly the upper bound.

As the argmax in (5) only copes with point estimates, these interval data should be processed by some other optimality criterion. We choose *interval-dominance*, *i.e.*, we reject an interval if its upper bound is lower than the lower bound of some other interval. Dominance should be checked for each pair of intervals; we end up with set $\mathcal{C}^* \subseteq \mathcal{C}$ of the classes associated to HMMs whose intervals are undominated, *i.e.*, $\mathcal{C}^*$ is the set:

$$\left\{ c_{d^*} \in \mathcal{C} \,\middle|\, \nexists d = 1, \ldots, D / \overline{P}(\mathbf{o}_1^{D+1}, \ldots, \mathbf{o}_T^{D+1} | \overline{\underline{\lambda}}_{d^*}) \leq \underline{P}(\mathbf{o}_1^{D+1}, \ldots, \mathbf{o}_T^{D+1} | \overline{\underline{\lambda}}_d) \right\}. \tag{6}$$

The above algorithm can be regarded as a *credal classifier* for time series, which may eventually return more than a single candidate class for the instance under consideration. An empirical validation of this approach is in the next section.

4. Experimental Validation on Human Action Recognition Tasks

In order to test our time series classifier, we consider a human action recognition problem (see Figure 4). For this specific application, the temporal data to be classified correspond to the features extracted by each frame from each video sequence. In particular we adopt the feature extraction algorithm proposed in [2], which describes the distribution of optical flows in the whole frame as an histogram with 16 bins representing directions. Flows are computed by block-matching in adjacent frames: this approximates instant velocities, and makes such information suitable for our approach. Simulation results on Weizmann [1] and KTH [5] benchmarks are reported in Tab. 1.

Notably, as a credal classifier is not returning a single class in output, the accuracy is not anymore a sufficient performances descriptor. For this reason we also consider: the percentage of instances classified with a single class (*determinacy*); the average number of classes returned when the classification is indeterminate (*indeterminate output size*); the accuracy over the instances classified with a single class (*single accuracy*); the accuracy over the instances classified with more classes (*set-accuracy*). As we also have data

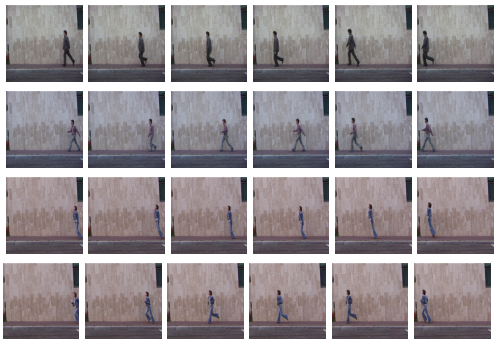

Figure 2. Human action recognition. Frames associated to different sequences depicting different actions.

about the precise classifier (*i.e.*, that obtained with the standard EM), we evaluate the accuracy of the precise method, when the imprecise is indeterminate (*credal-indeterminate precise accuracy*). The main comment to these results concerns this latter descriptor, whose low values show how the number of classes in output are an expressive indicator of the difficulty associated to the recognition of a particular action as depicted in a sequence.

	Weizmann		KTH	
Determinacy	84.72%	(61/72)	86.67%	(130/150)
Average Output Size	2.09	(23/11)	2.20	(44/20)
Single accuracy	70.49%	(43/61)	51.54%	(67/130)
Set-accuracy	54.55%	(6/11)	60.00%	(12/20)
Credal-Indeterminate Precise Accuracy	0.00%	(0/11)	20.00%	(4/20)

Table 1. Empirical validation of the action recognition algorithm proposed in Sect. 3. We performed leave-one-out classification with $N = 3$ for HMMs, and $s = 2$ for the imprecise EM.

5. Conclusions and Future Work

An imprecise-probabilistic version of the EM algorithm, specialised for HMMs has been proposed and adopted for likelihood-based time-series classification. The algorithm is effective in discriminating the hard-to-classify instances from the easy ones. If a single class-label is returned, this is very likely to be the correct one, while if multiple options are provided, the correct option is very likely to belong to this set. This approach seems to be particularly suited as a preprocessing tool for other classification algorithms. As a future work, we plan to extend to this framework other HMM-based techniques involving more sophisticated descriptors than the likelihood. In particular, some dissimilarity measures for HMMs (*e.g.*, [12]) could be adopted and hence extended to the imprecise-probabilistic case.

References

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *In IEEE Internation Conference on Computer Vision (ICCV)*, 2005.

[2] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[3] F. G. Cozman. Credal networks. *Artificial Intelligence*, 120:199–233, 2000.

[4] C. P. de Campos and F. G. Cozman. The inferential complexity of Bayesian and credal networks. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1313–1318, Edinburgh, 2005.

[5] I. Laptev and T. Lindeberg. Local descriptors for spatio-temporal recognition. In *Proc. of ICCV*, 2003.

[6] G. M. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, New York, 1997.

[7] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, 1988.

[8] L. Rabiner. A tutorial on HMM and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[9] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York, 1991.

[10] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *J. R. Statist. Soc. B*, 58(1):3–57, 1996.

[11] M. Zaffalon and E. Fagiuoli. Tree-based credal networks for classsification. *Reliable Computing*, 9(6):487–509, 2003.

[12] J. Zeng, J. Duan, and C. Wu. A new distance measure for hidden Markov models. *Expert Syst. Appl.*, 37:1550–1555, 2010.