Weighted maximum likelihood for parameters learning based on noisy labels in discrete Hidden Markov Models

P.J. Cano and E. Ramasso

FEMTO-ST institute, Dep. of Applied Mechanics & Dep. of Automatic Control and Micro-Mechatronic systems, Technopôle TEMIS, 25000 Besançon

Abstract. In supervised time-series segmentation, each instance in the training set has to be assigned a label. However, elicitation of labels from experts or their estimation may be time consuming and prone to errors. The problem considered in this paper is focused on time-series segmentation based on noisy and uncertain labels by using discrete Hidden Markov Models (dHMM). Maximum likelihood parameter learning in dHMM with such labels is tackled by two methods: the Evidential Expectation-Maximization (E2M) algorithm where weights represent plausibility functions, and the Weighted Likelihood Principle (WLP) coupled with the usual Expectation-Maximization algorithm. The model is tested using the E2M solution on simulated datasets. The results allows to evaluating the sensitivity of the quantization phase, with report to the noise level and the level of uncertainty on labels, on the quality of the statistical modelling of continuous-valued time-series.

1 Introduction

Hidden Markov models (HMM) are powerful tools for sequence modeling and state sequence recognition that have been used in many different applications. Discrete HMM represents a particular of HMM where the observations are discrete symbols. One of the most extended use has been text character recognition from several scripts as Latin [9], Korean [11] or Farsi (Arabic) [6]. Other applications concerned signal processing [16], video event classification [3] medical applications [1], model families of biological sequences [1] or transformer relaying protection [12].

A dHMM is composed of observed variables (outputs) $X_t, t = 1 \dots T$ where t is a discrete time index and latent discrete random variables (hidden states) Y_t [14]. The sequence of states $Y_1, Y_2, ...Y_T$ is a first-order Markov chain and the distribution of the output X_t at time t depends only on Y_t.

One of the objective of a dHMM is to estimate the state sequence hidden within the observations. In order to improve the convergence (quicker and more precise) and to better estimate the parameters, it is proposed to use partial prior knowledge about the states. For that, we first apply the Evidential Expectation-Maximization (E2M) algorithm [8] by assuming that the prior is encoded by a set of plausibility functions or basic belief assignments (Section 2). We then apply the Weighted Likelihood Principle (WLP) coupled with the Expectation-Maximization algorithm and we discuss the differences between both solutions. Experiments are focused on continuous-valued time-series segmentation with the solution provided by E2M. We illustrate the impact of the quantization phase with report to uncertain and noisy labels on the quality of the results (Section 3).

2 Developing the model

2.1 Model and notations

The following parameters are used to describe a HMM:

- Prior probabilities $\boldsymbol{\varPi} = \{\pi_1, ..., \pi_k, ..., \pi_K\}$, where $\pi_k = P(Y_1 = k)$ is the probability of being in state k at $t = 1$ being K the number of states;
- Transition probabilities $\mathbf{A} = [a_{kl}]$,where
$$a_{kl} = P(Y_t = l|Y_{t-1} = k), \quad (k, l) \in \{1, ..., K\}^2$$
is the probability for being in state l at time t given that it was in state k at $t - 1$ with $\sum_l a_{kl} = 1$;
- Observation symbol probabilities $\mathbf{B} = [b_{kv}]$ where
$$b_{kv} = P(x_t = v|Y_t = k), \;\; k \in \{1, ..., K\} \;\; \& \;\; v \in \{1, ..., V\}$$
is the probability for being in state k at time t and observing symbol v with $\sum_v b_{kv} = 1$

The set of parameters is denoted as $\theta = (\mathbf{A}, \mathbf{B}, \boldsymbol{\varPi})$.

The complete data is defined as $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ composed of the observed output sequence $\boldsymbol{x} = (x_1, ..., x_T)$ and the corresponding sequence of hidden states $\boldsymbol{y} = (y_1, ..., y_T)$. In the discrete case each observation takes a discrete value $v \in \{1, ..., V\}$ called symbol.

2.2 Learning procedures based on soft labels

E2M algorithm Let Y be a variable taking values in a finite domain $\varOmega = \{1, 2 \ldots K\}$, called the *frame of discernment*. Uncertain information about Y (i.e. partial knowledge about hidden states, also called soft labels) is supposed to be represented by a mass function m on $\varOmega$, $\sum_{A \subseteq \varOmega} m(A) = 1$ (assumed normalized).

Maximising the likelihood in presence of such uncertain information about hidden states can be performed by applying the E2M algorithm [8]. For that, it is first required to express the likelihood function over hidden and observed variables which, in the dHMM, is given by

$$\begin{aligned} L(\boldsymbol{\theta}; \boldsymbol{z}) &= p(y_1; \varPi) \Bigg(\prod_{t=2}^{T} p(y_t|y_{t-1}; \mathbf{A}) \Bigg) \prod_{t=1}^{T} p(x_t|y_t; \mathbf{B}) \\ &= \Bigg(\prod_{k=1}^{K} \pi_k^{y_{1k}} \Bigg) \Bigg(\prod_{t=2}^{T} \prod_{k,l} a_{kl}^{y_{(t-1,k)} y_{tl}} \Bigg) \Bigg(\prod_{t=1}^{T} \prod_{k=1}^{K} \prod_{v=1}^{V} b_{kv}^{y_{tk}} \Bigg) \end{aligned}$$

where y_{tk} is a binary variable such that $y_{tk} = 1$ if state k is true at time t. The second step is to take the conditional expectation of the log-likelihood given partial knowledge on states which can then be obtained at iteration q of E2M as [8]:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = \mathbb{E}_{\boldsymbol{\theta}^{(q)}}[\log(L(\boldsymbol{\theta}; \boldsymbol{z})|\boldsymbol{x}, pl] = \frac{\sum_{\boldsymbol{y} \in \Omega} \log(L(\boldsymbol{\theta}; \boldsymbol{z}))p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}^{(q)})pl(\boldsymbol{y})}{L(\boldsymbol{\theta}^{(q)}; \boldsymbol{x}, pl)}$$

where pl is the contour function (plausibility of singleton states) associated to m. $L(\boldsymbol{\theta}^{(q)}; \boldsymbol{x}, pl)$ is a generalized likelihood function [8] evaluated by using the forward-backward propagations [15]. By expanding the expectation, we get three terms:

- Two terms involving prior and transitions and similar to HMM with continuous observations [15] ;
- The third one is specific to the dHMM and concerns the emission probability model $\mathbf{B}$ from which the maximum likelihood estimate can be obtained as:

$$b_{kv}^{(q+1)} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{(q)} \, 1\{x_t = v\}}{\sum_{t=1}^{T} \gamma_{tk}^{(q)}}$$

where $\gamma_{tk} = \mathbb{E}_{\boldsymbol{\theta}^{(q)}}[y_{t,k}|\boldsymbol{x}, pl]$ has the same expression as in [15].

Weighted likelihood principle (WLP) It is described in detail in [18, 19] and aims at exploiting pieces of information obtained from independent samples generated by some distributions with unknown parameters that have justly to be estimated. In the WLP model, a sample is produced by a weighted likelihood function [13, 18]. For the dHMM, it is given by

$$L(\boldsymbol{\theta}; \boldsymbol{z}, \mathbf{W}) = p(y_1; \Pi)^{w_{1k}} \left(\prod_{t=2}^{T} p(y_t|y_{t-1}; A)^{w_{(t-1,k)} w_{tl}} \right) \prod_{t=1}^{T} p(x_t|y_t; B)^{w_{tk}}$$

which can be rewritten by using multinomial variables as

$$L(\boldsymbol{\theta}; \boldsymbol{z}, \mathbf{W}) = \left(\prod_{k=1}^{K} \pi_k^{w_{1k} y_{1k}} \right) \left(\prod_{t=2}^{T} \prod_{k,l} a_{kl}^{w_{(t-1,k)} y_{(t-1,k)} w_{tl} y_{tl}} \right) \left(\prod_{t=1}^{T} \prod_{k=1}^{K} \prod_{v=1}^{V} b_{kv}^{w_{tk} y_{tk}} \right) \quad (1)$$

where the weights $\mathbf{W} = \{w_{t,k}, t = 1 \dots T, k = 1 \dots K : w_{t,k} \geq 0\}$ can be obtained by optimization (given a target) [13, 19] or provided by an end-user. By taking the logarithm of Eq. 1, we have:

$$\log L(\boldsymbol{\theta}; \boldsymbol{z}, \mathbf{W}) = \sum_{k=1}^{K} w_{1k} y_{1k} \log \pi_k + \sum_{t=2}^{T} \sum_{k,l} w_{(t-1,k)} y_{(t-1,k)} w_{tl} y_{tl} \log a_{kl} + \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{v=1}^{V} w_{tk} y_{tk} \log b_{kv} \tag{2}$$

We then apply the usual EM algorithm [7] to estimate the parameters $\boldsymbol{\theta}$ in an iterative way as in standard dHMM. Assuming independence between hidden variables and weights, the expression of the expectation of $\mathbb{E}[w_{tk} y_{tk} | \boldsymbol{x}, \boldsymbol{\theta}]$ can be obtained as:

$$\mathbb{E}[w_{tk} y_{tk}] = \frac{w_{tk} p(y_t = k | \boldsymbol{x}, \boldsymbol{\theta})}{\sum_{l=1}^{K} w_{tl} p(y_t = l | \boldsymbol{x}, \boldsymbol{\theta})}$$

This posterior distribution is then used to find the expectation of the complete-data log likelihood evaluated for some general parameter value [2]. The M-step then makes use of this posterior that relies on soft labels to estimate the parameters for the next iteration.

Differences betweem the two models The E2M and WLP models differ from two main points, independently on the statistical model considered (dHMM or another).

Firstly, in E2M, the prior on latent variables is expressed as a plausibility function (in $[0, 1]$), while the WLP allows more general weights provided positiveness. In practice, it permits more flexibility. Real applications are necessary to assess if this difference actually plays a role, either for weights elicitation or estimation, or concerning the performance.

Secondly, and more fundamentally, the plausibilities used in E2M play a role of weights on the emission model that generates the likelihood of the current data given the current state ($p(x_t | y_t)$). Therefore, the computation of the posterior probability on states (γ_t) at time t makes use of the plausibilities at t (in the forward propagation [2]) and on $t + 1$ (in the backward propagation [2]). In comparison, in the WLP model, the weights are combined conjunctively only once with the posterior probability on states ($p(y_t | \boldsymbol{x})$). Eventually, this difference leads to models with different likelihoods, and more interestingly, it shows that the WLP acts similarly as the approach proposed in [4, 5].

3 Simulations

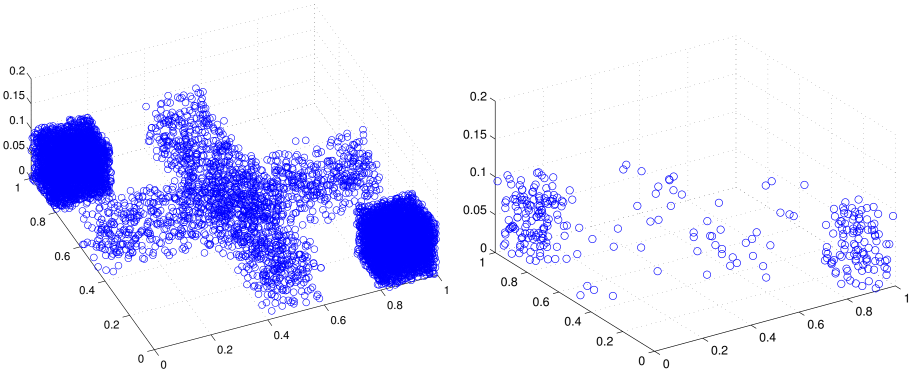
We consider a dHMM with 3 states and three symbols per state distributed with report to uniform distribution defined as:

$$\boldsymbol{\mathit{\Pi}} = (1/3, 1/3, 1/3)', \ \mathbf{A} = \begin{pmatrix} 0.6 \ 0.3 \ 0.1 \\ 0.1 \ 0.6 \ 0.3 \\ 0.1 \ 0.3 \ 0.6 \end{pmatrix}$$

$$S_1 \sim \begin{cases} x \sim \mathcal{U}(0, 0.2) \\ y \sim \mathcal{U}(0.8, 1) \\ z \sim \mathcal{U}(0, 0.1) \end{cases} \ S_2 \sim \begin{cases} x \sim \mathcal{U}(0.8, 1) \\ y \sim \mathcal{U}(0, 0.2) \\ z \sim \mathcal{U}(0, 0.1) \end{cases}$$

$$S_3 = \{S_{31}\} \cup \{S_{32}\} \quad S_{31} \sim \begin{cases} x \sim \mathcal{U}(0.4, 0.6) \\ y \sim \mathcal{U}(0, 1) \\ z \sim \mathcal{U}(0, 0.1) \end{cases} \quad S_{32} \sim \begin{cases} x \sim \mathcal{U}(0, 1) \\ y \sim \mathcal{U}(0.4, 0.6) \\ z \sim \mathcal{U}(0, 0.1) \end{cases}$$

Two sets of samples are represented in Figure 1.



(a) 10000 Points (b) 300 Points

Fig. 1: Distribution of signal points

The *Kmeans* algorithm was used for vector quantization [14] in order to transform those continuous-valued observations into discrete symbols. Different number of clusters were tested to estimate the impact of the quantization on the performance. Two different experiments were carried out with this model in order to study the influence of “label imprecision” and “labeling error” [5, 8, 15].

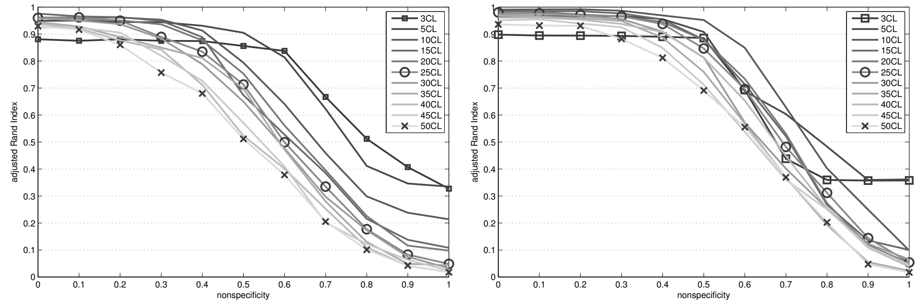
3.1 Influence of label imprecision

To study how the influence of imprecision of knowledge on hidden states affects the performing of the learning procedures described above, a learning sequence

$(\boldsymbol{x}, \boldsymbol{y})$ of length T was generated using the model above. Uncertain labels were generated as follows:

$$pl_{tk} \begin{cases} 1 & \text{if } y_t = k, \\ \nu & \text{otherwise.} \end{cases}$$

ν represents the nonspecificity coefficient, which quantifies the imprecision of the contour function pl_t. To assess the quality of learning, a testing dataset of 1000 observations was generated following the same distribution. The most probable state at a given time was given by the maximum a posteriori probability [14], assuming no previous knowledge about hidden states in the test sequence. The precision of the predicted state sequences was assessed using the adjusted Rand index (ARI) [10] (equals 0 on average for a random partition, and 1 when comparing two identical partitions). The whole experiment (data generation, clustering and learning) was repeated 30 times, for different number of clusters and for $T = 100$ and $T = 300$.



(a) T=100

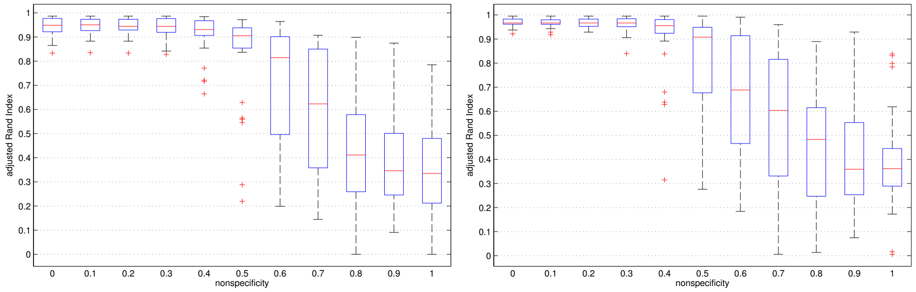
(b) T=300

Fig. 2: Medians of the adjusted Rand index as a function of the nonspecificity coefficient over 30 repetitions for diferent number of clusters, from 3 to 50.

Results are shown in Figures 2 and 3. We can observe that the results degrade from the fully supervised ($\nu = 0$) to the fully unsupervised ($\nu = 1$) case. In Figure 2 we can see different curves representing the results for different number of clusters. For a small number of clusters, the results with precise knowledge about states ($\nu < 0.4$) are lower than for a larger number of clusters. However, from that point and till the fully unsupervised situation, curves representing larger number of clusters decrease faster and reach values near to 0. Those with fewer number of clusters keep a higher ARI till $\nu = 1$ and do not decrease so fast.

3.2 Influence of labeling error

To simulate a situation where information on states may be wrong, we proceed as proposed in [5, 8, 15]. At each time step t, an error probability q_t was drawn



(a) T=100 (b) T=300

Fig. 3: Boxplots of the adjusted Rand index as a function of the nonspecificity coefficient over 30 repetitions for 5 clusters. Learning datasets of T=100 (left) and T=300 (right) observations.

randomly from a beta distribution with mean ρ and standard deviation 0.2. With probability q_t, the state y_t was then replaced by a completely random value $\tilde{\mathrm{y}}_t$ (with a uniform distribution over possible states). The plausibilities pl_{tk} were determined as

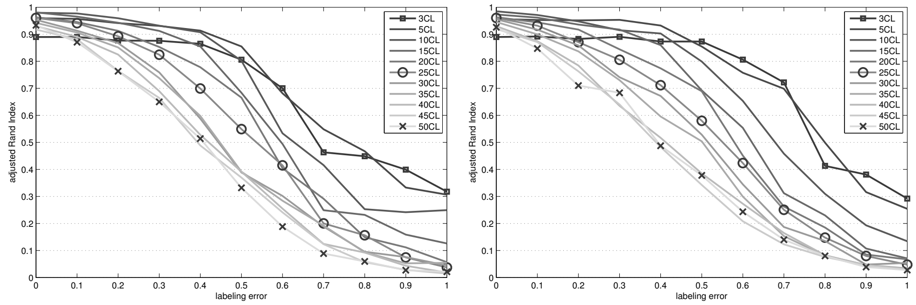
$$pl_{tk} = P(y_t = k|\tilde{y}_t) = \begin{cases} q_t/K + 1 - q_t & \text{if } \tilde{y}_t = k, \\ q_t/K & \text{otherwise.} \end{cases}$$

Uncertain labels are more imprecise when the error probability is high. Training and test data sets were generated as in previous section, and results were evaluated in the same way. For each randomly generated data set, the dHMM was applied with uncertain labels pl_{tk}, noisy labels $\tilde{y}_{tk}$ and no information on states.

Figure 4 shows the ARI as a function of mean error ρ for uncertain (left) and noisy (right) labels for different number of clusters and $T = 100$. As expected, a degradation of the segmentation quality is observed when the mean error probability ρ increases. The ARI tends to a value close to zero as ρ tends to 1 for a larger number of clusters. For fewer clusters, the results when ρ tends to 1 stay over 0. From the curves, we see that a smaller number of clusters give generally better results. The number of clusters used for quantization produces a side effect called distorsion [14] which remains difficult to assess in practice.

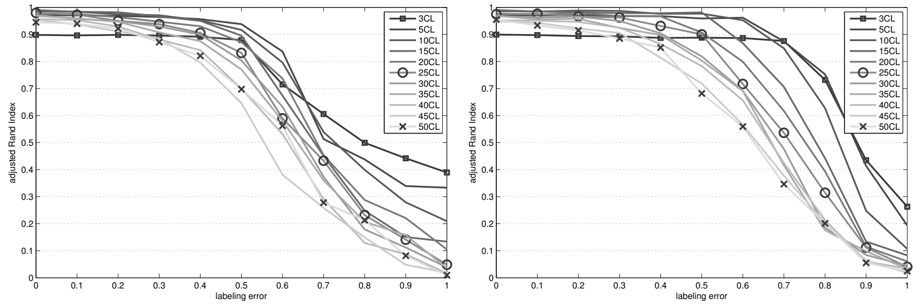
In Figure 5 we show the same experiments as in Figure 4 but with longer sequences ($T = 300$). Results are quite similar in both cases but we appreciate that with a larger number of observations, the curves scatter less and results are better for all values of ρ. This is an expected result since the dHMM is a statistical model where the parameters are learned by maximum likelihood and therefore the quantity of learning data may have an important impact on estimations.

Figure 6 shows the evolution of both the noisy and uncertain labels for the experiment with 5 clusters. It is proved that the use of partial information on states in the form of uncertain or noisy labels allows to reach better results than



(a) Uncertain labels, T=100 (b) Noisy labels, T=100

Fig. 4: Medians of the adjusted Rand index as a function of the labeling error for uncertain and noisy labels over 30 repetitions for different number of clusters, from 3 to 50. Learning datasets made of T=100 observations.



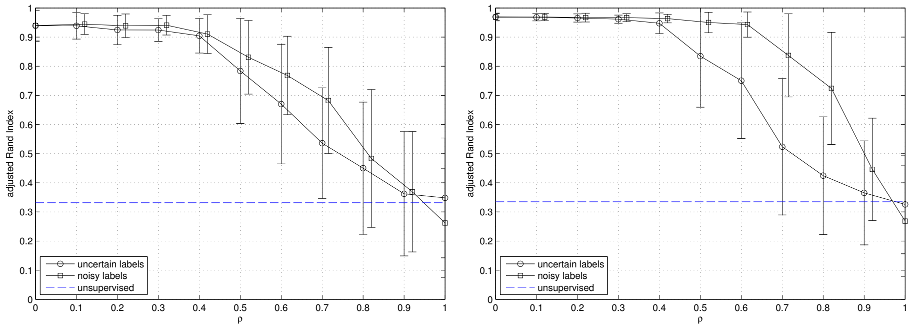
(a) Uncertain labels, T=100 (b) Noisy labels, T=300

Fig. 5: Medians of the adjusted Rand index as a function of the labeling error for uncertain and noisy labels over 30 repetitions for different number of clusters, from 3 to 50. Learning datasets made of T=300 observations.

the unsupervised case in every condition. Noisy labels reach better results than the uncertain labels till $\rho = 0.9$.

4 Conclusion

This paper studies the influence of labelling errors on the performance of of discrete Hidden Markov Models for continuous-valued time-series segmentation. Noisy and uncertain labels can be taken into account by the Evidential EM algorithm or by the weighted maximum likelihood principle, yielding two different results. The results shows that the degradation of the performance was accentuated when the quantization phase was inappropriately tuned. In contrast with the continuous HMM proposed in [15], the model can behave better when



(a) T=100

(b) T=300

Fig. 6: Average values (plus and minus one standard deviation) of the adjusted Rand index over the 30 repetitions, as a function of the mean error probability for learning datasets of T=100 (left) and T=300 (right) observations

considering noisy labels than uncertain labels. The way to integrate imprecise knowledge on latent variables in HMM is under study. This would lead to imprecise transition matrices and observation models generating sets of possible states sequences [17].

Acknowledgment

This work has been carried out in the framework of the Laboratory of Excellence ACTION through the program "Investments for the future" managed by the National Agency for Research (references ANR-11-LABX-01-01). The authors are grateful to the Région Franche-Comté and "Bpifrance financement" supporting the SMART COMPOSITES Project in the framework of FRI2.

References

1. Baldi, P., Chauvin, Y., Hunkapiller, T., McClure, M.A.: Hidden Markov models of biological primary sequence information. Proceedings of the National Academy of Sciences 91(3), 1059–1063 (1994)
2. Bishop, C.: Pattern recognition and machine learning. Springer (2006)
3. Chen, H.S., Tsai, W.J.: A framework for video event classification by modeling temporal context of multimodal features using HMM. Journal of Visual Communication and Image Representation 25(2), 285–295 (2014)
4. ming Cheung, Y.: A rival penalized em algorithm towards maximizing weighted likelihood for density mixture clustering with automatic model selection. In: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. pp. 633–636 (Aug 2004)
5. Côme, E., Oukhellou, L., Denoeux, T., Aknin, P.: Learning from partially supervised data using mixture models and belief functions. Pattern recognition 42(3), 334–348 (2009)

6. Dehghan, M., Faez, K., Ahmadi, M., Shridhar, M.: Handwritten Farsi (Arabic) word recognition: a holistic approach using discrete HMM. Pattern Recognition 34(5), 1057–1065 (2001)
7. Dempster, A.P.; Laird, N.R.D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39(1), 1–38 (1977)
8. Denoeux, T.: Maximum likelihood estimation from uncertain data in the belief function framework. Knowledge and Data Engineering, IEEE Transactions on 25(1), 119–130 (2013)
9. Elms, A., Procter, S., Illingworth, J.: The advantage of using an HMM-based approach for faxed word recognition. International Journal on Document Analysis and Recognition 1(1), 18–36 (1998)
10. Hubert, L., Arabie, P.: Comparing partitions. Journal of classification 2(1), 193–218 (1985)
11. Kim, H.J., Kim, S.K., Kim, K.H., Lee, J.K.: An HMM-based character recognition network using level building. Pattern recognition 30(3), 491–502 (1997)
12. Ma, X., Shi, J.: A new method for discrimination between fault and magnetizing inrush current using HMM. Electric Power Systems Research 56(1), 43–49 (2000)
13. Newton, M., Raftery, A.: Approximate bayesian inference with the weighted likelihood boostrap. Journal of the Royal Statistical Society 56, 3–48 (1994)
14. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
15. Ramasso, E., Denoeux, T.: Making use of partial knowledge about hidden states in HMMs: an approach based on belief functions. Fuzzy Systems, IEEE Transactions on 22(2), 395–405 (2014)
16. Ranjan, R., Mitra, D.: HMM modeling for OFDM–BER performance. AEU-International Journal of Electronics and Communications 69(1), 18–25 (2015)
17. Skulj, D.: Discrete time markov chains with interval probabilities. International Journal of Approximate Reasoning 50(8), 1314 – 1329 (2009)
18. Wang, S.: Maximum Weighted Likelihood Estimation. Ph.D. thesis, British Columbia (2001)
19. Wang, X., Zidek, J.: Derivation of mixture distributions and weighted likelihood function as minimizers of KL-divergence subject to constraints. Annals of the Institute of Statistical Mathematics 57(4), 687–701 (2005)