Sparseness Achievement in Hidden Markov Models

Manuele Bicego*
DEIR - University of Sassari
via Torre Tonda, 34
07100 Sassari - Italy

Marco Cristani
DI - University of Verona
Strada Le Grazie, 14
37134 Verona - Italy

Vittorio Murino
DI - University of Verona
Strada Le Grazie, 14
37134 Verona - Italy

Abstract

In this paper, a novel learning algorithm for Hidden Markov Models (HMMs) has been devised. The key issue is the achievement of a sparse model, i.e., a model in which all irrelevant parameters are set exactly to zero. Alternatively to standard Maximum Likelihood Estimation (Baum Welch training), in the proposed approach the parameters estimation problem is cast into a Bayesian framework, with the introduction of a negative Dirichlet prior, which strongly encourages sparseness of the model. A modified Expectation Maximization algorithm has been devised, able to determine a MAP (Maximum A Posteriori probability) estimate of HMM parameters in this Bayesian formulation. Theoretical considerations and experimental comparative evaluations on a 2D shape classification task contribute to validate the proposed technique.

1. Introduction

Hidden Markov Models (HMM - [21]) represent a probabilistic sequential tool widely employed in different research areas, such as speech recognition [21], handwritten character recognition [14], DNA and protein modelling [15], gesture recognition [10], behavior analysis and synthesis [17], 2D shape classification [13, 6] — only to cite a few.

A practical but fundamental issue to be addressed when using HMMs is the so called model selection, which regards the determination of the structure and the number of states of a model. The number of states typically represents a compromise between the goodness of the fitting to the data (better when increasing the number of states) and the generalization capability (ability to generalize to not previously seen situations). General approaches — such as BIC [23], MDL [22] or AIC [2] — as well as HMM specific approaches [27, 26, 16, 25, 7] have been proposed in the past to solve this problem. On the other side poor attention has been given to the determination of the structure of HMM, namely to the problem of choosing a particular topology, obtained by emphasizing or by disregarding the importance of some particular links between states in the model. In this context, some a priori determined topologies have been proposed with success, such as the well known left-to-right topology (firstly introduced in [4, 18]), or the circular structure proposed in the context of shape recognition [3]. More interesting are other data-driven approaches, aimed at extracting the structure of the HMM directly from the data [8, 9, 1].

In this paper a novel data-driven approach to the determination of the structure of an HMM is proposed, able to derive a *sparse* estimate of the model, which we called *Sparse Hidden Markov Model* (SHMM). In general, a model is said to be "sparse" when irrelevant or redundant components are *exactly zero*. Sparseness is highly desirable in supervised learning since it produces a structural simplification of the model, disregarding unimportant parameters: in this sense, a sparse model distills the information of all the training data providing only high representative parameters. Moreover, it has been shown in the context of the kernel-based methods that the generalization ability increases with the degree of sparseness, supporting the key idea behind Support Vector Machines [28]. Sparseness has been recently and successfully applied also in the context of Bayesian probabilistic supervised learning [11].

In the proposed approach, the sparseness of the model is induced by casting the HMM parameters estimation problem in a Bayesian framework with the introduction of a negative Dirichlet prior on the transition probabilities (with a flat prior on the other HMM parameters). Adopting a negative Dirichlet prior permits to discourage and penalize uniform distributions among state transition probabilities, leading to an annihilation process: low probabilities transitions are rapidly driven to zero, whereas strong ones are stirred up. The resulting model presents few very relevant parameters, whereas other irrelevant ones are exactly zero. Even if negative Dirichlet priors have been adopted in several contexts [12] — the well known Jeffrey's prior [5] is a Dirichlet prior itself — its employment in the HMM training, to the

14th International Conference on Image Analysis and Processing (ICIAP 2007)
0-7695-2877-5/07 $25.00 © 2007 IEEE

IEEE COMPUTER SOCIETY

best of our knowledge, has never been investigated. A modified Expectation-Maximization (EM) algorithm has been developed able to determine the MAP (Maximum A Posteriori) estimate of the HMM parameters.

The proposed method is similar in spirit to [8, 9], where a Bayesian approach has been proposed in order to determine the structure of HMM. Nevertheless, there are three main differences: the first and most important is that we used a Dirichlet prior, whereas in [8, 9] an entropic one has been used. The Dirichlet prior discourages uniform configurations more strongly than the entropic prior, favoring component annihilation more distinctly (see Fig. 1). Second, the parameters estimation derived from the application of the Dirichlet prior has a closed formed solution, whereas the estimation using the entropic prior does not. Finally, component annihilation is not fully automatic, but requires an additional test, while in the proposed approach an explicit and automatic rule is provided.

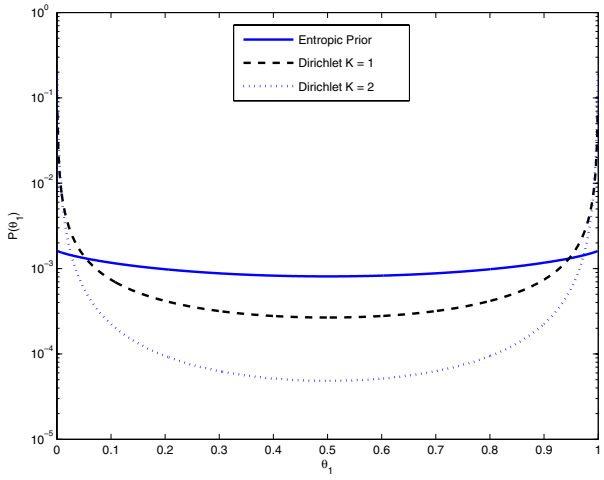

Figure 1. Plot of different priors for a two parameters configuration θ_1, θ_2. In this case, $\theta_2 = 1 - \theta_1$ and $P(\theta_1)$ is the binomial distribution, shown in a logarithmic axis. Please notice that all these priors discourage uniform configurations, but the Dirichlet ones operate more strongly.

Sparse HMMs have been tested in a real problem, involving 2D shape classification, and compared with standard ML estimate derived from Baum-Welch training. Obtained recognition results confirm the intuition that sparseness could lead to more robust estimation of HMM, able to generalize to unknown objects in a better fashion.

The rest of the paper is organized as follows. In Sect. 2 HMM are summarized, mainly to fix the notation. Sparse Hidden Markov Models are presented in Sect. 3, and an experimental validation is presented in Sect. 4. In Sect. 5, conclusions are drawn and future perspectives are envisaged.

2. Hidden Markov Models

A Hidden Markov Model [21] is a Markov Model where states are not directly observable, each state has associated a density function describing the probability of observing a certain symbol from that state. HMM is composed by the following entities: a set $S = \{S_1, S_2, \cdots, S_N\}$ of (hidden) states; a transition matrix $\mathbf{A} = \{a_{ij}\}$, where $a_{ij} \geq 0$ represents the probability of going from state S_i to state S_j; an emission matrix $\mathbf{B} = \{b(o|S_j)\}$, indicating the probability of emission of symbol o from state S_j; an initial state probability distribution $\boldsymbol{\pi} = \{\pi_i\}$, representing the probability of the first state $\pi_i = P[Q_1 = S_i]$.

For a sequence $\mathbf{O}$ and an HMM $\boldsymbol{\lambda}$, there is a standard recursive procedure able to compute the probability $P(\mathbf{O}|\boldsymbol{\lambda})$, and is called the *forward-backward* procedure [21].

2.1. HMM training

Given a sequence $\mathbf{O}$, there exists a well-established procedure able to determine the HMM parameters maximizing the probability $P(\mathbf{O}|\boldsymbol{\lambda})$. This technique, called the *Baum-Welch re-estimation* procedure [21], is an instance of the well-known *Expectation-Maximization* (EM) algorithm for Maximum Likelihood (ML) estimation. The E-step reduces to compute the following two variables, given the current model [21]:

- $\xi_t(i,j)$: it represents the probability of passing from state S_i at time t to state S_j at time $t+1$, given the observations and the model, *i.e.*

$$\xi_t(i,j) = P(Q_t = S_i, Q_{t+1} = S_j|\mathbf{O}, \boldsymbol{\lambda}) \qquad (1)$$

  Note that the sum of $\xi_t(i,j)$ over time t could be interpreted as the expected number of transitions from state S_i to state S_j.

- $\gamma_t(i)$: it represents the probability of being in state S_i at time t, given the observations and the model, *i.e.*

$$\gamma_i = P(Q_t = S_i|\mathbf{O}, \boldsymbol{\lambda}) \qquad (2)$$

Also in this case, the sum of $\gamma_t(i)$ over t can be interpreted as the expected number of transitions from S_i.

In the discrete case, i.e., when $b(o|S_i)$ is a discrete pdf, the M-step subsequently adjusts the model parameters on the basis of the computed variables:

$$\begin{aligned} \bar{\pi}_i &= \text{expected frequency in state } S_i \text{ at time } t = 1 \\ &= \gamma_1(i) \end{aligned} \qquad (3)$$

0-7695-2877-5/07 $25.00 © 2007 IEEE

IEEE COMPUTER SOCIETY

$$\bar{a}_{ij} = \frac{\text{exp. num. of trans. from } S_i \text{ to } S_j}{\text{exp. num. of trans. from } S_i} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (4)$$

$$\bar{b}(v_j|S_i) = \frac{\text{exp. num. of times in } S_i \text{ and obs. symb. } v_j}{\text{exp. num. of times in } S_i} = \frac{\sum_{t=1}^{T-1} \gamma_t(i) \text{ s.t. } O_t = v_j}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (5)$$

A similar strategy could be derived for continuous Gaussian HMMs [19, 20].

3. Sparse Hidden Markov Models

3.1. Dirichlet prior and posterior formulation

The sparseness of the model is induced by introducing a negative Dirichlet prior on the Bayesian estimation of the parameters. This prior, in the case of multinomial distributions (such as HMM transition probabilities) with N conditional probabilities $\boldsymbol{\theta} = \theta_1 \cdots \theta_N$, has the following simple and general form:

$$P_D(\boldsymbol{\theta}) \propto \boldsymbol{\theta}^{-K} = \prod_j \theta_j^{-K} \quad (6)$$

K drives the functional form of the prior. $K = 0.5$ reduces to the well known Jeffrey's prior [5].

Applying the prior in (6) to a multinomial yields the following posterior:

$$P(\boldsymbol{\theta}|\boldsymbol{\omega}) = \frac{P(\boldsymbol{\omega}|\boldsymbol{\theta})P_D(\boldsymbol{\theta})}{P(\boldsymbol{\omega})} \propto P(\boldsymbol{\omega}|\boldsymbol{\theta})P_D(\boldsymbol{\theta}) \propto \prod_j \theta_j^{\omega_j} \prod_j \theta_j^{-K} = \prod_j \theta_j^{\omega_j - K} \quad (7)$$

where ω_j is the evidence for the parameter θ_j.

3.2. MAP estimation

To derive the MAP estimation we solve the following constrained maximization

$$\theta_j^{MAP} = \max_{\theta_j} \mathcal{L}_{post} = \max_{\theta_j} \log(P(\boldsymbol{\theta}|\boldsymbol{\omega})) = \max_{\theta_j} \sum_i (\omega_j - K) \log(\theta_j) \quad (8)$$

subject to $\sum_j \theta_j = 1$. Setting the derivative to zero and applying the Lagrange multiplier ℓ

$$\frac{\partial}{\partial \theta_j} \left(\sum_j (\omega_j - K) \log(\theta_j) - \ell(\sum_j \theta_j - 1) \right) = 0 \quad (9)$$

we obtain

$$\theta_j^{MAP} = \frac{\omega_j - K}{\ell} \quad (10)$$

3.3. Modified Expectation-Maximization

Given the MAP solution, the following step is to cast it into the training scheme of the HMM. We have one set $\{\theta_{ij}\}$ for each state S_i, represented by the transition probabilities $\{a_{ij}\}$. The evidence $\boldsymbol{\omega}$ is determined in the E-step as usual, using the current model and computing the variables $\xi_t(i,j)$:

$$\omega_{ij} = \sum_{t=1}^{T-1} \xi_t(i,j) \quad (11)$$

The modified EM differs from the standard EM only in the M-step. More in details, the re-estimation formulas remain the same for the emission probabilities $b(o|S_i)$ and for the initial state probabilities $\boldsymbol{\pi}$, which is equivalent to the application of a flat prior. By applying the MAP solution in (10) the parameters a_{ij} are re-estimated as

$$a_{ij} = \frac{\max\left(\sum_{t=1}^{T-1} \xi_t(i,j) - K\ ,\ 0\right)}{\sum_{h=1}^{N} \max\left(\sum_{t=1}^{T-1} \xi_t(i,h) - K\ ,\ 0\right)} \quad (12)$$

Note that here the ℓ Lagrange parameter (see Eq.10) does disappear, due to the normalization process; moreover, the maximum is inserted in order to have all probabilities greater than zero. This formula, similar to the one proposed in [12], has a clear interpretation. All evidences are decreased by a factor K. This fact, together with the subsequent normalization process, increases even more high probabilities and decreases low probabilities, with a beneficial effect: probabilities which are poorly supported are more quickly driven to extinction, whereas strong probabilities are enforced. This is strengthened by the max operation: if a certain transition has not enough evidence it is

0-7695-2877-5/07 $25.00 © 2007 IEEE

IEEE COMPUTER SOCIETY

pruned from the model, forcing it to zero. This represents a clear and automatic way of pruning transitions.

Moreover, another pruning rule could in principle be derived, linked to the states: if, after the re-estimation step, a certain state has not active transitions (no transition incoming), then it could be pruned. So the proposed methodology could be also used in the context of number of states determination; we will investigate further this direction in the future.

4. Experimental session

Sparse HMMs have been compared against standard ML estimates in a 2D shape recognition problem [6]. Each object is represented by the sequence of the curvature coefficients, computed as follows: first, the contours are extracted by using the *Canny* edge detector; the boundary is then approximated by segments of approximately fixed length d_L. The resulting sequences show different lengths, ranging from 267 for the smallest object to 559 for the largest. Finally, the curvature value at point x is computed as the angle between the two consecutive segments intersecting at x. The initial point is the rightmost point lying on the horizontal line passing through the object centroid, following the boundary in a counterclockwise manner. A thorough analysis of the HMMs' capabilities in classifying 2D shapes is presented in [6], where the standard Maximum Likelihood method was tested in cases of translation, rotation, noise, occlusions, shearing transformations, and combination of the above perturbations, showing really satisfying results.

In this paper, we compare Sparse HMM with a version of the system described in [6]: unlike in that paper, the number of states was fixed to three for all experiments. Testing was performed on part of the object set used in [24], composed by seven classes, each containing 12 different shapes (shown in Fig. 2). Accuracies are computed using the Leave One Out scheme.

Regarding Sparse HMM, the number of states was again fixed to three: different values of the parameter K in the Dirichlet prior have been tested. For all the values, the training process converges usually after few iterations, maximizing the posterior of the model. In some cases, the posterior behavior is not strictly monotonic; actually, after the trimming operation of a transition parameter (see Eq.12), the posterior exhibits an abrupt downward fluctuation, that however does not corrupt the general posterior increasing behavior. This fluctuating trend is similar to that shown in [12]; anyway, it will be subject of future study. In order to understand the sparsification process, in Fig. 3 two matrices of transitions are displayed, relative to two three states HMM trained on the first shape. The matrices are derived from a standard training (left) and from the proposed training (right)–with K= 0.5– both starting from the same

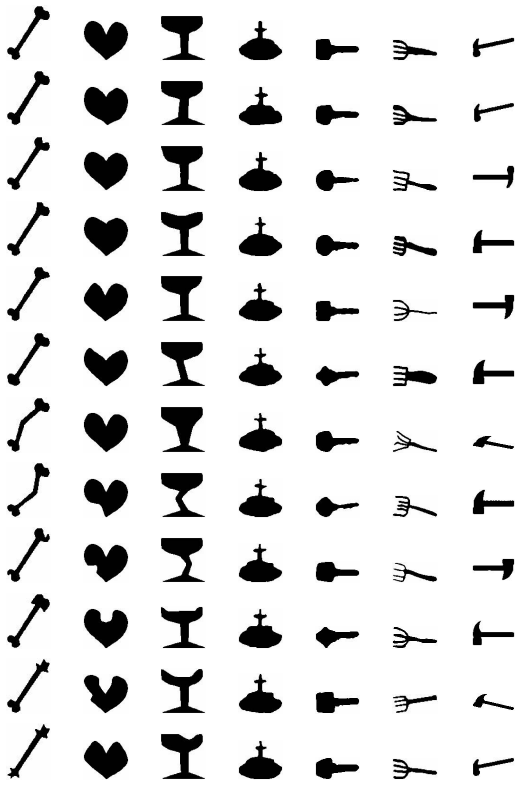

Figure 2. Objects set used for testing.

initialization. It is clear how the sparsification works: high values on the left matrix are enforced in the sparse version, whereas low values are weakened. It is also interesting to note that two transitions are exactly zero in the sparse transition matrix.

The results of the evaluation of our method are displayed in Table 1: from the table it is evident the gain in the performances gathered by the achievement of the model's sparseness: sparse HMMs outperform the standard rule for all values of parameter. In particular, looking at the results above, one can notice that the maximum of the performances is reached for $K = 0.5$. When increasing even more the

0.9539	0.0277	0.0184
0.0561	0.9439	0.0000
0.0112	0.0000	0.9888

0.9625	0.0234	0.0140
0.0497	0.9503	0
0.0057	0	0.9943

Figure 3. Transition matrices for a three state HMM trained on the first shape: on the left there is the matrix derived from the standard training, on the right the one derived from the sparse proposed method.

0-7695-2877-5/07 $25.00 © 2007 IEEE

IEEE COMPUTER SOCIETY

Method	Accuracy
Standard ML	81.43%
SHMM $K = 0.25$	91.43%
SHMM $K = 0.5$	92.50%
SHMM $K = 0.75$	90.83%
SHMM $K = 1.0$	90.12%
SHMM $K = 1.25$	88.69%
SHMM $K = 1.5$	89.64%
SHMM $K = 1.75$	87.86%

Table 1. Accuracies of SHMMs on 2D shape recognition experiment for different values of K.

parameter value, the beneficial effect of sparseness seems to decrease, even if remaining significantly higher than the Maximum Likelihood estimate result.

5. Conclusions

In this paper, a novel learning algorithm for Hidden Markov Model is proposed. The training problem is cast into a Bayesian framework by the introduction of a negative Dirichlet prior, which aims at individuating a model in which the transition table is sparse, i.e. peaked around few values that consistently represent the state transition behavior of the observed data. A modified EM algorithm has been proposed, able to derive a MAP estimate of the parameters. Experimentally, we proved that the method provides models that outperform ordinarily learned HMMs, with respect to shape classification tasks. Further studies inspired by our approach are currently under work.

References

[1] K. Abou-Moustafa, M. Cheriet, and C. Suen. On the structure of hidden markov models. *Pattern Recognition Letters*, 25:923–931, 2004.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, AC-19:716–723, 1974.

[3] N. Arica and F. Yarman-Vural. A shape descriptor based on circular Hidden Markov Model. In *IEEE Proc. Int Conf. Pattern Recognition*, volume 1, pages 924–927, 2000.

[4] R. Bakis. Continuous speech word recognition via centisecond acoustic states. In *Proc. of ASA Meeting*, 1976.

[5] J. Bernardo and A. Smith. *Bayesian Theory*. John Wiley & Sons, Chichester (UK), 1994.

[6] M. Bicego and V. Murino. Investigating Hidden Markov Models' capabilities in 2D shape classification. *IEEE Trans. on Pattern Analysis and Machine Intelligence - PAMI*, 26(2):281–286, 2004.

[7] M. Bicego, V. Murino, and M. Figueiredo. A sequential pruning strategy for the selection of the number of states in Hidden Markov Models. *Pattern Recognition Letters*, 24(9–10):1395–1407, 2003.

[8] M. Brand. An entropic estimator for structure discovery. In M. Kearns, S. Solla, and D. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11. MIT Press, 1999.

[9] M. Brand. Structure learning in conditional probability models via an entropic prior and parameter extinction. *Neural Computation*, 11:1144–1182, 1999.

[10] S. Eickeler, A. Kosmala, and G. Rigoll. Hidden Markov Model based continuous online gesture recognition. In *IEEE Proc. Int. Conf. Pattern Recognition*, volume 2, pages 1206–1208, 1998.

[11] M. Figueiredo. Adaptive sparseness for supervised learning. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(9):1150–1159, 2003.

[12] M. Figueiredo and A. Jain. Unsupervised learning of finite mixture models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.

[13] Y. He and A. Kundu. 2-D shape classification using Hidden Markov Model. *IEEE Trans. Pattern Analysis Machine Intelligence*, 13(11):1172–1184, 1991.

[14] J. Hu, M. Brown, and W. Turin. HMM based online handwriting recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(10):1039–1045, 1996.

[15] R. Hughey and A. Krogh. Hidden Markov Model for sequence analysis: extension and analysis of the basic method. *Comp. Appl. in the Biosciences*, 12:95–107, 1996.

[16] S. Ikeda. Construction of phoneme models - model search of hidden Markov models-. In *Proc. Int. Workshop on Intelligent Signal Processing and Communication Systems*, pages 82–87, 1993.

[17] T. Jebara and A. Pentland. Action reaction learning: Automatic visual analysis and synthesis of interactive behavior. In *Proc. Int Conf. Computer Vision Systems*, 1999.

[18] F. Jelinek. Continuous speech recognition by statistical methods. *Proc. of IEEE*, 64:532–536, 1976.

[19] B. Juang, S. Levinson, and M. Sondhi. Maximum likelihood estimation for multivariate mixture observations of Markov Chain. *IEEE Trans. Information Theory*, 32(2):307–309, 1986.

[20] L. Liporace. Maximum lilelihood estimation for multivariate observations of Markov sources. *IEEE Trans. on Information Theory*, IT-28(5):729–734, 1982.

[21] L. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. of IEEE*, 77(2):257–286, 1989.

[22] J. Rissanen. Stochastic complexity and modeling. *The Annals of Statistics*, 14:1080–1100, 1986.

[23] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[24] T. Sebastian, P. Klein, and B. Kimia. Recognition of shapes by editing Shock Graphs. In *Proc. Int Conf. Computer Vision*, pages 755–762, 2001.

[25] H. Singer and M. Ostendorf. Maximum likelihood successive state splitting. In *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 2, pages 601–604, 1996.

0-7695-2877-5/07 $25.00 © 2007 IEEE

IEEE COMPUTER SOCIETY

[26] A. Stolcke and S. Omohundro. Hidden Markov Model induction by Bayesian model merging. In S. Hanson, J. Cowan, and C. Giles, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 11–18. Morgan Kaufmann, San Mateo, CA, 1993.

[27] J. Takami and S. Sagayama. A successive state splitting algorithm for efficient allophone modeling. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 1, pages 573–576, 1992.

[28] V. Vapnik. *Statistical Learning Theory*. John Wiley, New York, 1998.

IEEE COMPUTER SOCIETY