Observable Markov Models

Oleg Golovko, Alexei Piskunov
e-mail: agp1@smtp.ru

May 2, 2005

Summary. *This article introduces an observable model equivalent to Hidden Markov Models. The model does not contain hidden part and has same major properties as respective HMM. OMMs also direct to the noncritical and obvious improvements in the major algorithms.*

1 Introduction

We will use following notions.

Definition 1. *Oriented graph G is the pair (EG, VG), where $EG \subset VG \times VG$. VG is called set of vertices and EG is called set of edges. For any edge $x = (\mu, \nu)$ denote $\mu = dom\,(x)$, $\nu = cod\,(x)$.*

Definition 2. $I = [0, 1]$.

Definition 3. *Probability space is the triple (Ω, F, P), where $F \subseteq 2^{\Omega}$ is σ - algebra of subsets of Ω and $\mathrm{P} : F \to \mathbb{R}$ is probability measure on F ([2]).*

Definition 4. *If Ω either $\mathbb{R}$ or countable then P is called probability distribution on Ω ([3]). If $\Omega = \mathbb{R}$ then F is Borel algebra on $\mathbb{R}$, which is unique, if Ω is countable then $F = 2^{\Omega}$ and therefore definition of Ω in these cases uniquely determines F and we will say that P is probability distribution on Ω.*

Definition 5. *Let A be graph called primary graph and B be set of letters called alphabet or observables and $\forall \mu \in VA$ there are probability distributions b_{μ} on B and a_{μ} on $\{(\mu, \nu) \in EA\}$. b_{μ} is called distribution of observable for the vertex μ and a_{μ} is called transition distribution from the vertex μ. These two sets of distributions could be interpreted as functions $b : VA \times B \to I$ and $a : VA \times EA \to I$ respectively. Function $\iota : VA \to \mathbb{R}$- probability distribution on VA*

called initial state distribution. Following [1] under HMM we will understand $\lambda = (a, b, \iota)$.

In voice recognition systems it's usually assumed that for the primary graph $VA = \{\nu_0, \nu_1, \nu_2, \nu_3, \nu_4\}$ and

$$EA = \{(\nu_0, \nu_1), (\nu_1, \nu_1), (\nu_1, \nu_2), (\nu_2, \nu_2), (\nu_2, \nu_3), (\nu_3, \nu_3), (\nu_3, \nu_4)\},$$

i.e. A is "chain" with loops on internal vertices and $\iota(\nu) = \left\{ \begin{array}{l} 1, \nu = \nu_0 \\ 0, \nu \neq \nu_0 \end{array} \right.$.

1.1 Hidden Markov Models (HMM), basic algorithms

There are 3 problems of HMM, but we will be interested in #1 and #3 only. Let $\overline{T} = \{0, ..., T-1\}$ and $\overline{T-1} = \{0, ..., T-2\}$, $O_T = \left\{o_t \in B | t \in \overline{T}\right\}$ - observation sequence of letters of alphabet B, $Q_T = \left\{q_t \in VA | t \in \overline{T}\right\}$ - sequence of the states.

Problem #1. Compute $\mathrm{P}(O_T|\lambda)$, the probability that given model λ generates given sequence O_T.

Problem #3. Adjust model λ to maximize $\mathrm{P}(O_T|\lambda)$ for the given O_T.

Algorithms. For the solution of the problem #1 forward and backward algorithms could be used, each requires $O\left(|VA|^2 \cdot T\right)$ operations.

Forward algorithm. Denote forward variable $\alpha_t(\mu) = \mathrm{P}(o_0, ..., o_t | q_t = \mu, \lambda)$, i.e. the probability of observation of the sequence $\{o_0, ..., o_t\}$ and at the moment t state is μ.

Solution. *1)* $\forall \mu \in VA : \alpha_0(\mu) = \iota(\mu) \cdot b_\mu(o_0)$

2) $\forall t \in \overline{T-1}$, $\forall \mu \in VA :$

$$\alpha_{t+1}(\mu) = \left(\sum_{\nu \in VA \wedge (\nu, \mu) \in EA} \alpha_t(\nu) \cdot a(\nu, \mu) \right) \cdot b_\mu(o_{t+1})$$

3) $P(O_T|\lambda) = \sum\limits_{\mu \in VA} \alpha_{T-1}(\mu)$

Backward algorithm. Denote backward variable $\beta_t(\mu) = \mathrm{P}(o_{t+1}, ..., o_{T-1} | q_t = \mu, \lambda)$, i.e. the probability of observation of the sequence $\{o_{t+1}, ..., o_{T-1}\}$ and at the moment t state is μ.

Solution. *1)* $\forall \mu \in VA : \beta_{T-1}(\mu) = 1$

2) $\forall t \in \overline{T-1}$, $\forall \mu \in VA :$

$$\beta_t(\mu) = \sum_{\nu \in VA \wedge (\mu, \nu) \in EA} \beta_{t+1}(\nu) \cdot a(\mu, \nu) \cdot b_\nu(o_{t+1})$$

3) $P(O_T|\lambda) = \sum\limits_{\mu \in VA} \iota(\mu) \cdot b_\nu(o_0) \cdot \beta_0(\mu)$

Solution of the problem #3, Baum-Welsh algorithm. Let $\forall t \in \overline{T-1}$, $\forall (\mu,\nu) \in EA$:

$$\begin{aligned}\xi_t(\mu,\nu) &= \mathrm{P}\left((q_t,q_{t+1}) = (\mu,\nu)\,|O_T,\lambda\right) \\ &= \frac{\alpha_t(\mu)\cdot a(\mu,\nu)\cdot b_\nu(o_{t+1})\cdot \beta_{t+1}(\nu)}{\mathrm{P}(O_T|\lambda)} \\ &= \frac{\alpha_t(\mu)\cdot a(\mu,\nu)\cdot b_\nu(o_{t+1})\cdot \beta_{t+1}(\nu)}{\sum\limits_{(\rho,\sigma)\in EA}\alpha_t(\rho)\cdot a(\rho,\sigma)\cdot b_\sigma(o_{t+1})\cdot \beta_{t+1}(\sigma)}\end{aligned}$$

Let $\gamma_t(\mu) \equiv \mathrm{P}(q_t = \mu|O_T,\lambda) = \sum\limits_{\nu\in VA\wedge(\mu,\nu)\in EA} \xi_t(\mu,\nu)$

$\sum\limits_{t\in\overline{T-1}} \gamma_t(\mu)$ - expected number of transitions from the vertex μ,

$\sum\limits_{t\in\overline{T-1}} \xi_t(x)$ - expected number of transitions over the edge x.

Most difficult part of the solution is minimization problem based on the parameters:

$\overline{\pi}_\mu = \gamma_0(\mu)$

$\overline{a}(\mu,\nu) = \frac{\sum\limits_{t\in\overline{T-1}}\xi_t(\mu,\nu)}{\sum\limits_{t\in\overline{T-1}}\gamma_t(\mu)}$

$\overline{b}_\mu(c) = \frac{\sum\limits_{t\in\overline{T-1}\wedge o_t=c}\gamma_t(\mu)}{\sum\limits_{t\in\overline{T-1}}\gamma_t(\mu)}.$

It could be done in many ways, but exact algorithms are not important for us now.

2 Observable Markov Models (OMM)

Definition 6. *For given HMM $\lambda = (a,b,\iota)$ define model $OMM(\lambda) = (\widetilde{a},\widetilde{\iota})$ with the graph $\widetilde{A}$ and functions $\widetilde{\iota}$ and π_B. $\widetilde{A}$ consists of vertices $V\widetilde{A} = VA\times B$, called states, and edges*

$$E\widetilde{A} = \{(f,c,d)\,|f\in EA\wedge c,d\in B\} = EA\times B\times B,$$

$\widetilde{a} : E\widetilde{A}\to I$ is defined as $\widetilde{a}(f,c,d) = b_{cod(f)}(d)\cdot a(f)$, $\widetilde{\iota} : V\widetilde{A}\to I$ is defined as $\widetilde{\iota}(\mu,c) = \iota(\mu)\cdot b_\mu(c)$ and $\pi_B : V\widetilde{A}\to B$ is projection.

Obviously graph $\widetilde{A}$ is more complex then primary graph A, but general model simplifies, because it consists of single probability distribution now instead of two ones and single set of vertices instead of two different sets of vertices and observables.

Forward algorithm for OMM. Denote forward variable $\alpha_t\left(\mu,o_t\right) = \mathrm{P}\left(o_0,...,o_t\right|\left(q_t,o_t\right)=\left(\mu,o_t\right),OMM\left(\lambda\right))$, i.e. the probability of observation $\{o_0,...,o_t\}$ and at the moment t state is $\left(\mu,o_t\right)$.

Solution. *1)* $\forall\left(\mu,o_0\right)\in V\widetilde{A}:\alpha_0\left(\mu,o_0\right)=\widetilde{\iota}\left(\mu,o_0\right)$
2) $\forall t\in\overline{T-1}$, $\forall\left(\mu,o_t\right)\in V\widetilde{A}:$

$$\begin{aligned}\alpha_{t+1}\left(\mu,o_{t+1}\right) &= \sum_{\nu\in VA\wedge(\nu,\mu)\in EA}\alpha_t\left(\nu,o_t\right)\cdot a\left(\nu,\mu\right)\cdot b_\mu\left(o_{t+1}\right)\\ &= \sum_{\nu\in VA\wedge((\nu,\mu),o_{t-1},o_t)\in E\widetilde{A}}\alpha_t\left(\nu,o_t\right)\cdot\widetilde{a}\left(\left(\nu,\mu\right),o_t,o_{t+1}\right)\end{aligned}$$

Consider $\left(\left(\mu,\nu\right),o_t,o_{t+1}\right)=\left(x,y\right)\in E\widetilde{A}$ *it could be rewritten as:*

$$\alpha_{t+1}\left(x\right)=\sum_{y\in V\widetilde{A}\wedge(y,x)\in E\widetilde{A}}\alpha_t\left(y\right)\cdot\widetilde{a}\left(y,x\right)$$

3) $P\left(O_T|OMM\left(\lambda\right)\right)=\sum\limits_{\mu\in VA}\alpha_{T-1}\left(\mu,o_{T-1}\right)\equiv\sum\limits_{(\mu,o_{T-1})\in V\widetilde{A}}\alpha_{T-1}\left(\mu,o_{T-1}\right)$

Backward algorithm for OMM. Denote backward variable $\beta_t\left(\mu,o_t\right) = \mathrm{P}\left(o_{t+1},...,o_{T-1}\right|\left(q_t,o_t\right)=\left(\mu,o_t\right),OMM\left(\lambda\right))$, i.e. the probability of observation $\{o_{t+1},...,o_{T-1}\}$ and at the moment t state is $\left(\mu,o_t\right)$.

Solution. *1)* $\forall\left(\mu,o_{T-1}\right)\in V\widetilde{A}:\beta_{T-1}\left(\mu,o_{T-1}\right)=1$
2) $\forall t\in\overline{T-1}$, $\forall\left(\mu,o_t\right)\in V\widetilde{A}:$

$$\begin{aligned}\beta_t\left(\mu,o_t\right) &= \sum_{\nu\in VA\wedge(\mu,\nu)\in EA}\beta_{t+1}\left(\nu,o_t\right)\cdot a\left(\mu,\nu\right)\cdot b_\nu\left(o_{t+1}\right)\\ &= \sum_{\nu\in VA\wedge((\mu,\nu),o_t,o_{t+1})\in E\widetilde{A}}\beta_{t+1}\left(\nu,o_{t+1}\right)\cdot\widetilde{a}\left(\left(\mu,\nu\right),o_t,o_{t+1}\right)\end{aligned}$$

Consider $\left(\left(\mu,\nu\right),o_t,o_{t+1}\right)=\left(x,y\right)\in EM_1$ *it could be rewritten as:*

$$\beta_t\left(x\right)=\sum_{y\in V\widetilde{A}\wedge(x,y)\in E\widetilde{A}}\beta_{t+1}\left(y\right)\cdot\widetilde{a}\left(x,y\right)$$

3) $P\left(O_T|OMM\left(\lambda\right)\right)=\sum\limits_{\mu\in VA}\widetilde{\iota}\left(\mu,o_0\right)\cdot\beta_0\left(\mu\right)\equiv\sum\limits_{(\mu,o_0)\in V\widetilde{A}}\widetilde{\iota}\left(\mu,o_0\right)\cdot\beta_0\left(\mu,o_0\right)$

Solution of the problem #3, Baum-Welsh algorithm for OMM.

Let $\forall t\in\overline{T-1}$, $\forall\left(\left(\mu,\nu\right),o_t,o_{t-1}\right)\in E\widetilde{A}$:

$$\begin{aligned}
\xi_t\left(\left(\mu,\nu\right),o_t,o_{t+1}\right) &= \mathrm{P}\left(\left(q_t,q_{t+1}\right)=\left(\mu,\nu\right)|O_T,\lambda\right)\\
&= \frac{\alpha_t\left(\mu,o_t\right)\cdot a\left(\mu,\nu\right)\cdot b_\nu\left(o_{t+1}\right)\cdot\beta_{t+1}\left(\nu,o_{t+1}\right)}{\sum\limits_{(\rho,\sigma)\in EA}\alpha_t\left(\rho,o_t\right)\cdot a\left(\rho,\sigma\right)\cdot b_\sigma\left(o_{t+1}\right)\cdot\beta_{t+1}\left(\sigma,o_{t+1}\right)}\\
&= \frac{\alpha_t\left(\mu,o_t\right)\cdot\widetilde{a}\left(\left(\mu,\nu\right),o_t,o_{t+1}\right)\cdot\beta_{t+1}\left(\nu,o_{t+1}\right)}{\sum\limits_{\substack{\left(y,z\right)\in E\widetilde{A}\\ \wedge\pi_B\left(y\right)=o_t\wedge\pi_B\left(z\right)=o_{t+1}}}\alpha_t\left(y\right)\cdot\widetilde{a}\left(y,z\right)\cdot\beta_{t+1}\left(z\right)}
\end{aligned}$$

Consider $\left(\left(\mu,\nu\right),o_t,o_{t+1}\right)=\left(x,y\right)\in E\widetilde{A}$ and in the sum in the denominator: $\pi_B\left(y\right)=o_t$ and $\pi_B\left(z\right)=o_{t+1}$, it could be rewritten as:

$$\xi_t\left(x,y\right)=\frac{\alpha_t\left(x\right)\cdot\widetilde{a}\left(x,y\right)\cdot\beta_t\left(y\right)}{\sum\limits_{(y,z)\in E\widetilde{A}}\alpha_{t-1}\left(y\right)\cdot\widetilde{a}\left(y,z\right)\cdot\beta_t\left(z\right)}$$

Construction of OMM proves the following statement.

Statement 1. *For each HMM λ there is exists $OMM\left(\lambda\right)$ such that* $\mathrm{P}\left(O_T|\lambda\right)=\mathrm{P}\left(O_T|OMM\left(\lambda\right)\right)$.

This means that HMM does not contain any additional features and it hidden part could be taken into account by construction of respective OMM wich is completely observable.

Conclusion. *We have defined the model that is equivalent to the HMM, but contains no hidden part and single probability distribution instead of two independent ones appeared in HMM. This allows to redefine HMM in more consistent mathematical way and to improve major HMM algorithms, though these improvements are so trivial that likely has been taken into account in all implementations anyway. At the same time proposed model by itself has become less intuitive and has lost it direct connection to the problems that have derived it.*

Acknowledgement. *We are grateful to William Mutual the chairman of UniSay for the opportunity to complete this paper. The result has been obtained during the design and development of the multimedia processing and speech timing system for the UniSay [5].*

References

[1] Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, vol.77, No.2, February 1989, http://www.ai.mit.edu/~murphyk/Bayes/rabiner.pdf

[2] http://www.en.wikipedia.org/wiki/Probability_theory

[3] http://www.en.wikipedia.org/wiki/Probability_distribution

[4] Rakesh Dugad, U. B. Desai ”A Tutorial on Hidden Markov Models”, http://www.uirvli.ai.uiuc.edu/dugad/hmm_tut.html

[5] http://www.unisay.com