Hollmén,J. and V.Tresp (2000). A hidden markov model for metric and event-based data.In Proceedings of EUSIPCO 2000 - X European Signal Processing Conference ,Volume II,pp.737 –740.

A HIDDEN MARKOV MODEL FOR METRIC AND EVENT-BASED DATA

Jaakko Hollmén† *and Volker Tresp*‡
† Helsinki University of Technology, Laboratory of Computer and Information Science
P. O. Box 5400, 02015 HUT, Finland, e-mail:Jaakko.Hollmen@hut.fi
‡ Siemens AG, Corporate Technology, Information and Communications
81730 Munich, Germany, e-mail:Volker.Tresp@mchp.siemens.de

ABSTRACT

The question of data representation is central to any data analysis problem. Ideally, the representation should faithfully describe the domain to be analyzed and in addition, the model used should be able to process such a representation. In practice, however, the modeler must often compromise how the problem is described, since the class of possible representations is constrained by the model. This problem may be circumvented by extending conventional models to handle more unconventional data representations. These data are often found in industrial environments and especially in telecommunications. In this paper, we consider an extension of hidden Markov models (HMM) for modeling data streams, which switch between metric and event-based representations. In a HMM, the representation of the observed data is constrained by the emission probability density. Since this density can not change its representation once it is fixed, modeling data streams involving different types of data semantics can be difficult. In the extension introduced in this paper, an additional data semantics variable is introduced, which is conditional on the hidden variable. Furthermore, data itself is conditioned on its semantics, which enables correct interpretation of the observed data. We briefly review the essentials of HMMs and present our extended architecture. We proceed by introducing inference and learning rules for the extension. As an application, we present a HMM for user profiling in mobile communications networks, where the data exhibits switching behavior.

1 INTRODUCTION

Hidden Markov models (HMM) are widely used in sequence processing and speech recognition [1, 8, 10, 7, 2]. The key benefit of HMM is their ability to model temporal statistics of data by introducing a discrete hidden variable that undergoes a transition from one time step to the next according to a stochastic transition matrix. At each time step, the HMM emits symbols, whose statistics are dependent on the current state of the hidden variable. Distribution of the emission symbols is embodied in the assumption of the emission probability density. Since this density can not change representation between different data semantics, data streams which involve different types of data can be difficult to model.

In this paper, we extend the HMM to handle cases where the data stream may switch between metric- and event-based representations. We define events to be categorical data with discrete outcomes, which are represented in discrete-time. The extension presented in this paper involves an introduction of a variable *data semantics*. This variable indicates whether data is to be interpreted as metric- or event-based. Data is conditioned on the data semantics variable in order to ensure correct interpretation. Moreover, the data semantics variable is dependent on the hidden state variable. This has importance in user profiling, where hidden states are thought to form a subpopulation, and where the data semantics becomes a quantity of its own importance.

Data exhibiting switching behavior is prevalent in industrial environments and telecommunications. It may rise as an inherent data generating mechanism of the industrial process, or may be introduced to the task by a feature extraction step. In an industrial process, an event may for example signify a condition under which a metric measurement can not be made. Feature extraction in turn aggregates a set of data points by calculating statistics over the set. These statistics are not always defined, as in the case of an empty set. This setting leads to special situations, which require the notion of an event variable. In the experiments, we use call data to model user behavior in a mobile communications network. We use average length of calls per day as our feature variable, and in case of no calls, we introduce an event *no calling*, since the average value is not defined on an empty set. This event is considered an important part of the user profiles describing user's behavior.

In Section 2, we briefly review the basic concepts of a HMM. In Section 3, we present the extended architecture allowing the semantics of data to change. The inference and learning procedures are presented in the framework of maximum likelihood estimation. In Sec-

tion 4, experiments in a user profiling problem are presented. The paper ends with Summary in Section 5.

2 HIDDEN MARKOV MODEL (HMM)

A hidden Markov model [7, 10, 2] assumes a discrete hidden variable s_t, which can be in one of exhaustive and mutually exclusive n states. The state changes in time stochastically according to a transition matrix $A = (a_{ji}) = P(s_t = j|s_{t-1} = i); i, j = 1, \ldots, n$. The density of the observed variable y_t depends on the state of the hidden variable and is defined by $P(y_t|s_t = j; \theta_1)$, parameterized bt θ_1. Denoting the hidden variables by $S = \{s_0, \ldots, s_T\}$, the observed variables by $Y = \{y_0, \ldots, y_T\}$, and the prior of the state and data at $t = 0$ by $P(s_0, y_0)$, the joint probability density parameterized by a_{ji} and θ_1 becomes

$$P(S, Y) = P(s_0, y_0) \prod_{t=1}^{T} P(s_t|s_{t-1}; A) \prod_{t=1}^{T} P(y_t|s_t; \theta_1).$$

Observed variables y_t may be either continuous or discrete, which is determined by the choice of a distribution. Discrete observations are easily modeled by assuming multinomial distributions or count distributions where appropriate; continuous measurements call for parametric distributions like the normal distribution, finite mixture distributions [5], or neural networks [3]. Inference and learning in the framework of maximum likelihood estimation are solved with the EM algorithm [1, 4, 9].

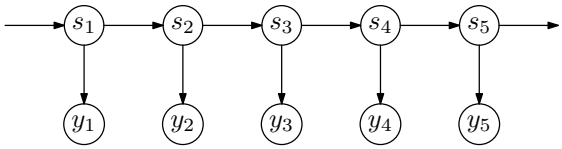

Figure 1: A HMM assumes a hidden variable s_t that changes state stochastically in time, the observations y_t are dependent on the hidden variable s_t through a emission probability density $P(y_t|s_t)$.

3 HMM FOR METRIC AND EVENT-BASED DATA

The standard HMM is defined by a single emission probability density, which models either continuous or discrete variables. We are interested in formulating an alternative model, where data can be either continuous (metric) data or discrete (event-based) data without restricting ourselves to one kind. There are many applications, where a more faithful problem representation is achieved by considering switching between event-based and metric data. We define events to be categorical data with discrete outcomes represented in discrete-time. An example of the representation is presented in Table 1.

Time index:	1	2	3	4	5	6
Time series data:	3.4	5.6	a	3.4	a	b
transformed data y_t:	3.4	5.6	1	3.4	1	2
semantics y_t^*:	m	m	e	m	e	e

Table 1: Original, switching time-series transformed to separate time-series for data and data semantics. Events a and b must be enumerated and have been assigned numerical values 1 and 2, respectively.

The solution involves conditioning the emission probability density of data on the new variable *data semantics*, which enables the correct interpretation of data. Furthermore, the data semantics variable is dependent on the hidden variable s_t. Denoting the data by y_t and data semantics by y_t^*, and all of the data semantics $Y^* = \{y_0^*, \ldots, y_T^*\}$, the joint probability of the variables in the model is

$$P(S, Y, Y^*) = P(s_0, y_0, y_0^*) \prod_{t=1}^{T} P(s_t|s_{t-1}; A)$$

$$\times \prod_{t=1}^{T} P(y_t|s_t, y_t^*; \theta^2) \prod_{t=1}^{T} P(y_t^*|s_t; \theta^3).$$

By conditioning on the semantics of the data one chooses the right type of model to be used in calculating the probability of data, which in turn is used in calculating the posterior of the hidden state. Note also that the variable data semantics is dependent on the state, which reflects the frequencies of each type of data in each state. In the case of metric variables, we have for the data semantics $y_t^* = m$ and when we encounter event based data, we have $y_t^* = e$. We are interested in estimating

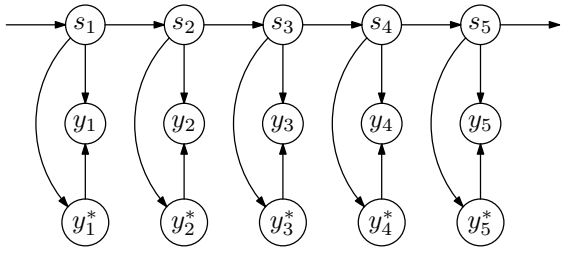

Figure 2: The extended HMM introduces a data semantics variable y_t^* that enables the correct interpretation of the observed variable y_t and is conditioned on the hidden state. The variable y_t^* is assumed to be observed.

the parameters of this model from data and once the model is learned, we want to estimate the state of the hidden variable based on observed variables (inference). The learning is done in the framework of maximum likelihood estimation using the EM algorithm [1, 4, 9].

3.1 Inference: Filtering and Smoothing

Inference is the procedure of estimating the state of the hidden variables given the observed variables. Inference can be divided into two parts: filtering and smoothing. In filtering, we move forward in time and infer on the states of the hidden variables given data up to present time. In smoothing, we move backward in time and infer on the hidden variables given all data available. The extended model is no longer a tree structured model as is the HMM, but a directed acyclic graph. The inference can be implemented using the junction tree algorithm [6, 11] for the underlying graphical model. We can write the equations for inference as iterative update rules for the model. The forward part of the inference involves calculating the predicted state of the hidden variable s_t given observed variables up to time $t-1$ and the density of s_{t-1}. This is based on the transition matrix $A = (a_{ji})$. The notation Y_t means that we have data available up to time t.

$$P(s_t = j|Y_{t-1}, Y^*_{t-1}) = \sum_{i=1}^{n} a_{ji} P(s_{t-1} = i|Y_{t-1}, Y^*_{t-1}).$$

The filtered state of the hidden variable is calculated based on the predicted state, the observed data and its semantics. In effect, the predicted state is multiplied by the joint probability of the data and the semantics

$$P(s_t = j|Y_t, Y^*_t) = c \cdot P(s_t = j|Y_{t-1}, Y^*_{t-1}) \times P(y^*_t|s_t = j) P(y_t|s_t = j, y^*_t).$$

c is a normalizing constant. The forward part of the inference consist of applying these two equations for $t = 1, \ldots, T$, after which the values of $P(s_t|Y_t, Y^*_t)$ are estimated. The key to the extended architecture is the explicit decoupling of the data semantics and the data itself, which is reflected in the inference procedure in the decoupling of the joint probability density of data and its semantics according to the conditional independence assumptions in the model. Conditional independence assertion allows a factorization of the two, which enables using each dimension independently in a user profiling problem.

In smoothing, the information flows backward in time and the states of the hidden variables are estimated given all data. This is important in retrospective analysis of time series, and is also needed in learning. Smoothing consists of two equations, which are evaluated alternately for $t = T-1, \ldots, 1$. These are equivalent in the standard formulation of the hidden Markov model,

$$P(s_t = j, s_{t+1} = i|Y_T, Y^*_T) = P(s_t = j|Y_t, Y^*_t) \frac{P(s_{t+1} = i|Y_T, Y^*_T)}{P(s_{t+1} = i|Y_t, Y^*_t)} a_{ij}.$$

By marginalizing with regard to s_{t+1}, we get the posterior of s_t given all data

$$P(s_t = j|Y_T, Y^*_T) = \sum_{i=1}^{n} P(s_t = j, s_{t+1} = i|Y_T, Y^*_T).$$

3.2 Learning

In order to learn the parameters from data, we apply the EM algorithm, which is a iterative algorithm consisting of an E-step which is implemented through the inference procedure described in the last section and the maximization step, where the parameters are updated accordingly. The M-step for the transition probabilities in the Markov chain is the same as in the standard HMM

$$a^{(new)}_{ji} = \frac{\sum_{t=1}^{T} P(s_t = j, s_{t-1} = i|Y_T, Y^*_T)}{\sum_{t=1}^{T} P(s_{t-1} = i|Y_T, Y^*_T)}.$$

It is natural to consider events $e_i, i = 1, \ldots, k$ to be multinomially distributed with probability of event i occurring in the state j as $P(y_t = i|y^*_t = \mathrm{e}, s_t = j) = \theta^2_{ij}$ with the constraint that $\sum_i \theta_{ij} = 1$. For the state dependent event probabilities we have the following M-step

$$\theta^{2,(new)}_{ij} = \frac{\sum_{t=1, y^*_t = e, y_t = i}^{T} P(s_t = j|Y_T, Y^*_T)}{\sum_{t=1, y^*_t = e}^{T} P(s_t = j|Y_T, Y^*_T)}.$$

The priors for the data semantics can easily be calculated, since the semantics are always observed

$$\theta^{3,(new)}_{j} = \frac{\sum_{t=1, y^*_t = e}^{T} P(s_t = j|Y_T, Y^*_T)}{\sum_{t=1}^{T} P(s_t = j|Y_T, Y^*_T)}.$$

The probability for data semantics to be metric is then $P(y^*_t = m|s_t = j) = 1 - P(y^*_t = e|s_t = j) = 1 - \theta^3_j$, since the semantics can be one of the two. The M-step for the observed metric data model is dependent on the kind of model used. Interesting alternatives are the Gaussian or exponential distributions (length distribution) or mixtures thereof.

4 EXPERIMENTS

In user profiling, one is interested in expressing expected behavior of users through models. In mobile communications networks, call detail records store attributes of individual calls and the call data collectively describes the calling behavior of mobile phone subscribers. In the data, calling activity of fraudulent and normal users were recorded. We use a two-state hidden Markov model for learning two user profiles from this call data. The time index t in our HMM denotes a day. For the observed variable we have the average length of the calls per day; if no calls are made, there is an event *no calling*. The variable data semantics is set accordingly. For instance, if we have an average call length of 5.5 minutes, we have $y_t = 5.5$ and $y^*_t = m$, since the data has a metric representation. If we have an event, we have $y_t = 1$ and $y^*_t = e$. Notice that events are enumerated. Representing average of zero calls by zero or by missing data would make modeling of such data possible with a standard HMM, but would be simply incorrect. Moreover, the statistics about different semantics occurring forms an independent part of its own

in our user profiling formulation. We used call data from 200 mobile phone users in estimating the model parameters. The switching time series describing the calling behavior were of variable lengths. Altogether there were 11388 recorded calling days. The EM algorithm was performed for 20 iterations. Transition matrix was given to reflect the dynamic characteristics of the states ($a_{11} = 0.93$, $a_{22} = 0.7$). Our model iden-

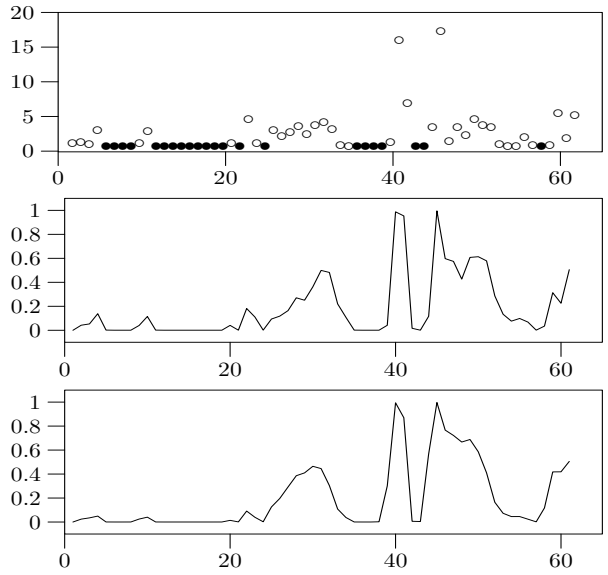

Figure 3: In the top panel, data is shown. Events *no calling* are marked with black circles, metric measurements are marked with open circles. Time-varying posteriors for the state $s_t = 2$ are shown in the lower panels. The middle panel shows the filtered $P(s_t = 2|Y_t, Y_t^*)$ probabilities, the bottom panel shows the smoothed $P(s_t = 2|Y_T, Y_T^*)$ probabilities.

tified the first state as the infrequent user with lots of events *no calling* $P(y_t^* = e|s_t = 1) = 0.56$. In the other state, average length of calls was larger and the event no calling very rare $P(y_t^* = e|s_t = 2) = 0.004$. In the state $s_t = 1$, the identified average length of the calls was 1.66 minutes, in state $s_t = 2$ the average length was 8.35 minutes. Exponential distribution for the call lengths was assumed.

5 SUMMARY

In data analysis problems, data representation acts as a mediator between the problem domain and the model. Ideally, it should describe the world well and be suitable for the model. Often, representation is compromised to be compatible with the model. An alternative approach is to develop extensions of established models in order to preserve the data representation that best serves the purpose without having to compromise how the problem is described. In this paper, we extended the HMM for modeling time-series that exhibit switching between metric- and event-based representations.This essentially combines an HMM with continuous emission distribution and one with discrete emission distribution. An additional variable data semantics controls the interpretation of data and is dependent on the hidden variable. Inference and learning rules were developed within a maximum likelihood framework. The approach was illustrated in a user profiling problem, where the mechanism leading to the event representation was important from user profiling point of view. In further work, the ideas concerning the conditioning of the data on its semantics is generalized to Bayes networks or in the context of neural networks.

References

[1] Leonard E. Baum. An inequality and associated maximation technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3:1–8, 1972.

[2] Yoshua Bengio. Markovian models for sequential data. *Neural Computing Surveys*, 2:129–162, 1999.

[3] Chris Bishop. *Neural Networks in Pattern Recognition*. Oxford Press, 1996.

[4] A. P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[5] B.S. Everitt and D.J. Hand. *Finite Mixture Distributions*. Monographs on Applied Probability and Statistics. Chapman and Hall, 1981.

[6] Finn V. Jensen. *An Introduction to Bayesian Networks*. UCL Press, 1996.

[7] B.H. Juang and L.R. Rabiner. Hidden markov models for speech recognition. *Technometrics*, 33(3):251–272, 1991.

[8] S.E. Levinson, L.R. Rabiner, and M.M. Sondhi. An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. *The Bell System Technical Journal*, 62(4):1035–1074, 1983.

[9] Geoffrey J. McLahlan. *The EM Algorithm and Extensions*. Wiley & Sons, 1996.

[10] Alan B. Poritz. Hidden markov models: A guided tour. In *Proceedings of the IEEE International conference of Acoustics, Speech and Signal Processing (ICASSP'88)*, pages 7–13, 1988.

[11] Padhraic Smyth, David Heckerman, and Michael I. Jordan. Probabilistic independence networks for hidden markov probability models. *Neural Computation*, 9(2):227–269, 1997.