Dynamic Non-Parametric Mixture Models and The Recurrent Chinese Restaurant Process

Amr Ahmed AMAHMED@CS.CMU.EDU
Eric Xing EPXING@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University

1. The Temporal Dirichlet Process Mixture Model

Dirichlet process mixture models provide a flexible Bayesian framework for estimating a distribution as an infinite mixture of simpler distributions that could identify latent classes in the data [1]. However the full exchangeability assumption they employ makes them an unappealing choice for modeling longitudinal data such as text, audio and video streams that can arrive or accumulate as epochs, where data points inside the same epoch can be assumed to be fully exchangeable, whereas across the epochs both the structure (i.e., the number of mixture components) and the parameterizations of the data distributions can evolve and therefore unexchangeable.

To solve this problem, we present the temporal Dirichlet process mixture model (TDPM) as a framework for modeling these complex longitudinal data, in which the number of mixture components at each time point is unbounded; the components themselves can retain, die out or emerge over time; and the actual parameterization of each component can also evolve over time in a Markovian fashion. For lack of space, we restrict ourselves here to one construction of our process that we call the Recurrent Chinese Restaurant Process (RCRP), for other two equivalent constructions, please refer to [2].

1.1. Recurrent Chinese Restaurant Process

The RCRP, showin in Figure 1, is a generalization of the CRP. The RCRP operates in epochs, say, days. Customers entered the restaurant in a given day are not allowed to stay beyond the end of this day. At the end of each day, the consumptions of dishes are analyzed by the owner of the restaurant who assumes that popular dishes will remain popular in the next day, and uses this fact to plan the ingredients to be bought, and the seating plan for the next day. To encourage customers in the next day to try out those pre-planed dishes, he records on each table the dish which was served there, as well as the number of customers who shared it. As another incentive, he allows the first customer to set on such a table to order a variation of the dish recorded there. The generative process proceeds as follows. At day t, customer i can pick an empty table, k, that was used to serve dish $\phi_{k,t-1}$, and had $n_{k,t-1}$ customers, with probability equals to $\frac{n_{k,t-1}}{N_{t-1}+i+\alpha-1}$, he then chooses the current flavor of the dish,$\phi_{k,t}$, distributed according to $\phi_{k,t} \sim P(.|\phi_{k,t-1})$. If this retained table k has already $n_{k,t}^{(i)}$customers[1], then he joins them with probability $\frac{n_{k,t-1}+n_{k,t}^{(i)}}{N_{t-1}+i+\alpha-1}$ and shares the current flavor of the dish there. Alternatively, he can pick a *new empty* table that was not used in the previous day,$t-1$, with probability $\frac{\alpha}{N_{t-1}+i+\alpha-1}$, and orders a new dish, K^+, $\phi_{K^+,t} \sim G_0$. Finally, he can share a *new* table k, with $n_{k,t}^{(i)}$ customers, with probability $\frac{n_{k,t}^{(i)}}{N_{t-1}+i+\alpha-1}$ and shares the newly ordered dish with them.

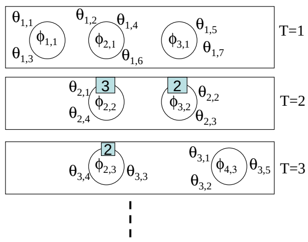

Figure 1. The recurrent Chinese restaurant process, see [2] for other two constructions.

1.2. Modeling Higher Order Dependencies

To allow for higher order dependencies between the time-dependent processes, we incorporated historic usage patterns by *decaying* their contributions exponentially over time epochs. A similar idea has been proposed in [3], however in [3], each epoch has exactly one data point.We define two new hyperparmaters, kernel width, λ, and history size, W. The owner of the restaurant now records on each table, not only its us-

[1]The superscript (i), indicates the same quantity just before customer i.

age pattern on day $t-1$, but its weighted cumulative usage pattern over the last W days. Where the weight associated with the count from day $t-h$ is given by $\exp^{\frac{-h}{\lambda}}$, and as such the contribution from epoch $t-h$ decays exponentially over time. A customer $x_{t,n}$ entering the restaurant at time t will behave exactly in the same way as before using the new numbers recorded on the table. This makes the cost of running one Gibbs iteration is $O(n \times W)$, moreover, an active mixture is considered dead if and only if, it is not used for exactly W contiguous echoes. It is interesting to note that these two new hyper-parameters allow the TDPM to degenerate to either a set of independent DPMs at each epoch when W=0, and to a global DPM, i.e ignoring time, when $W = T$ and $\lambda = \infty$. In between, the values of these two parameters affect the expected life span of a given chain. In fact in [2], we found that the chain durations followed a power law distribution.

1.3. Illustrations

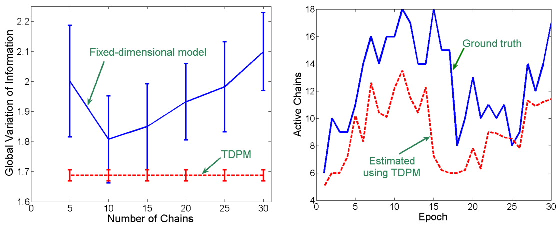

Figure 2. Illustrating results on simulated data. Left panel contrast the accuracy of the recovered clustering, using global variation of information to fixed models, right panel illustrates the TDPM ability to vary the number of clusters/chains over time

We show two applications of our framework, first we built an infinite dynamic mixture of Gaussian factors (infinite variable-duration mixture of Kalman filters).We let each chain represent the evolution of the mean parameter of a Gaussian distribution with a fixed covariance. The chain dynamics is taken to be a linear state-space model reduced to a random walk for simplicity. We simulated 100 date points from this model over 30 epochs. We ran Gibbs sampling for 1000 iterations and then took 10 samples every 100 iterations for evaluations. We compare our findings in Figure 2 to the ground truth, and to that produced from a fixed dynamic mixture of Kalman Filters with various number of chains.

Fixing the above model but changing the emission model into a logistic-normal one, we obtain a simple dynamic topic model in which each document belong to only one topic, as opposed to a parametric LDA-based one as in [4] . We apply this model over the NIPS12 collection with results shown in Figure 3.

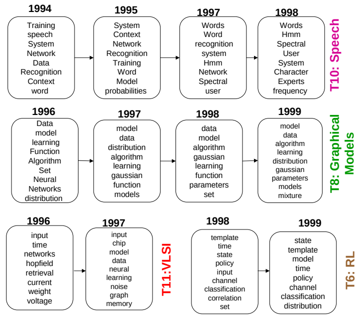

Figure 3. Illustrating results on the NIPS12 dataset: note that the topics have variable durations and start/end times. See [2] for more illustrations.

1.4. Related and Future Work

Several approaches have been recently proposed to solve the same fundamental problem addressed in this paper. With the exception of [3,5], most approaches use the stick-breaking construction to couple the weights and/or atoms of nearby DPMs as in ordered-based DPMs[6]. However, we believe that utilizing the CRP directly is easier to implement and enables us to model the rich-gets-richer phenomenon, which captures a wide range of applications. Recently, [5] showed a similar construction to the one presented here using the CRP in which customers are stochastically allowed to leave the restaurant at each epoch, therefore, the customer's chance of persisting in the restaurant decays exponentially over time. We captured this effect *deterministically* via the time-decaying kernel we introduced earlier. It still remains as an open problem to *objectively* contrast these approaches and possibly unify them. Moreover, we are currently building a full-fledged non-parametric dynamic topic model by extending the simple construction we showed in Figure 3 by replacing the emission model to be a DP for each document.

1.5. Acknowledgment
This work was supported by NIH 1090151 and NSF IIS-0713379 awards.

References

[1] C. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. The Annals of Statistics, 2(6):1152-1174, November.

[2] A. Ahmed and E. Xing. Dynamic Non-Parametric Mixture Models and the Recurrent Chinese Restaurant Process, (SDM2008).

[3] X. Zhu, Z. Ghahramani, and J. Lafferty. Time-sensitive Dirichlet process mixture models. Technical Report CMU-CALD-05-104, 2005.

[4] D. Blei and J. Lafferty. Dynamic topic models. ICML 2006.

[5] F. Caron, M. Davy,A. Doucet. Generalized Polya urn for time-varying Dirichlet processes, UAI 2007.

[6] J.Griffn and M. Steel. Order-based dependent Dirichlet processes. Journal of the American Statistical Association,101(473)