Estimating the Parameters of Randomly Interleaved Markov Models

Daniel Gillblad, Rebecca Steinert (Authors)
Swedish Institute of Computer Science
Box 1263, SE-164 29 Kista, Sweden
Emails: {dgi, rebste}@sics.se

Diogo R. Ferreira (Author)
IST – Technical University of Lisbon
Campus Taguspark, Portugal
Email: diogo.ferreira@ist.utl.pt

***Abstract*—Sequences that can be assumed to have been generated by a number of Markov models, whose outputs are randomly interleaved but where the actual sources are hidden, occur in a number of practical situations where data is captured as an unlabeled stream of events. We present a practical method for estimating model parameters on large data sets under the assumption that all sources are identical. Results on representative examples are presented, together with a discussion on the accuracy and performance of the proposed estimation algorithms. Finally, we describe a real-world case study where we apply the technique to the sequence of events recorded in the technical support database of an IT vendor.**

I. Introduction

In many situations, data generated by a number of distinct processes can only be observed as a single sequence of interleaved events, with no information on which source a particular event originated from. Such sequences may occur in a variety of logs, such as records over financial transactions, business processes, or production system events in industrial plants. To be able to understand the processes that generate this type of data, we need to be able to estimate the parameters of a model representing this situation from example sequences.

We present a statistical model consisting of a mixture of Markov chains for describing this type of interleaved sequence data, and we focus on the specific problem of how to estimate the parameters of such model.

Directly related model descriptions are not common. A recent work by [1] addresses a similar kind of model with a focus on activity recognition, where the model is first trained with complete data and then applied to a test set of incomplete data. Here we aim at estimating the model parameters from an single, unlabeled stream of data. In [2], the authors discuss the problem of inferring mixtures of Markov chains and the complexity thereof by incrementally dividing an observed sequence into disjoint subsets. Although similar model assumptions are made, the authors propose very different algorithms for parameter estimation compared to the ones we discuss here.

Somewhat analogous to the model we describe here are Hidden Markov Models (HMMs), which are commonly used to model sequential patterns [3]. However, a HMM represents a stochastic process generated by an underlying Markov chain that is observed through a distribution of the possible output states. These are model assumptions that are indeed rather different from the randomly interleaved mixture of Markov models studied here. More similar is the hierarchical hidden Markov model (HHMM) [4]. It generalizes the HMM by representing each of the hidden states with a HHMM in itself, making most states output sequences rather than single symbols. Another relevant extension to the HMM is the factorial hidden Markov model [5], which uses multiple independent chains of hidden variables, and the distribution of the observed variable is conditional on the states of all hidden variables at the same time step. However, none of these models effectively express the interleaved nature of our problem.

In applications such as user click-stream analysis [6], [7] and recent applications in economics such as wage mobility [8], the term *mixture of Markov models* has been used in connection with the problem of sequence clustering, where the goal is to classify a set of given sequences into different groups. Also, the mover-stayer model – with a wide range of applications from economics [9], [10] to health care [11], [12] and for which an EM algorithm has been devised [13] – is sometimes referred to as a mixture, while it can be regarded as an extended Markov model.

By *mixture of Markov models* we understand, in this paper, the result of interleaving events coming from several unknown Markov sources. The goal is to estimate the parameters of the sources in terms of *selection probabilities* and *transition probabilities*, when the source that produced each event is unknown. We study a special case of mixtures of Markov chains, namely when all sources have identical transition probability matrices. In section II we introduce the problem and notation, and in section III describes how estimate the model parameters. Section IV presents test results on both synthetic models and a real-world application.

II. Problem statement

In the general case, let us assume that there is a set of K distinct sources, and that the behavior of each source can be captured by a Markov chain. Each of these Markov chains $M_1, \dots, M_K$ is time-homogeneous, ergodic, and has a finite state space. Let A_k where $k = 1, 2, \dots, K$ be the state transition matrix for each source, where an element $A_k(a, b)$ represents the probability of source k moving from state a to state b, or $A_k(a, b) = P(b \mid a; M_k)$. Let each state be assigned a different symbol, such as “A”, “B”, “C”, etc. Each source produces an output when it changes to a new state; the output is the symbol that corresponds to that state. A series of transitions will result in an output stream of symbols containing the states that the source went through in a certain period of time. The current state of the source at the start of the observation is unknown.

The system is such that the output of all sources is written to a single stream, where the symbols produced by each source become interleaved with symbols coming from other sources. Let the output stream be denoted by the sequence $\boldsymbol{x} = \{x_1, \ldots, x_N\}$ with length N, where each symbol x_n comes from one of the K sources. The process of mixing the outputs of all K sources into a single output sequence $\boldsymbol{x}$ is modeled as a stationary source selection process with probabilities $\{\pi_1, \pi_2, \ldots, \pi_K\}$, as shown in Fig. 1. At each step n it is as if one source k is being selected with probability $\pi_k(n) = P(s_n = k) = \pi_k$, which is independent of n. The observable sequence $\boldsymbol{x}$ is then created by randomly selecting a source, drawing the next symbol from this source, and repeating this process for each subsequent position.

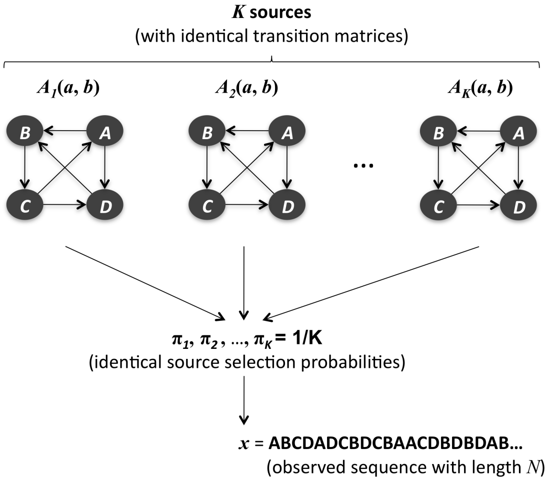

K sources
(with identical transition matrices)
$A_1(a, b)$
$A_2(a, b)$
$A_K(a, b)$
B
A
C
D
...
$\pi_1, \pi_2, \ldots, \pi_K$ = 1/K
(identical source selection probabilities)
x = ABCDADCBDCBAACDBDBDAB...
(observed sequence with length *N*)

Fig. 1. The model consists of K identical Markovian sources and a source selection mechanism. In the observed output sequence, the source that produced each symbol is unknown.

The only observable output from this system is the symbol sequence $\boldsymbol{x}$. Here, we will study the special case of identical sources with $A_k(a, b) = A(a, b)$ and equal source selection probabilities $\pi_k = \pi = 1/K$. Under these assumptions, all model parameters can be efficiently estimated directly from a symbol sequence.

III. Mixtures of identical Markov models

A. Estimating the source parameters

To estimate the parameters of the Markov chains $A_k(x_i, x_j) = A(x_i, x_j)$ in a mixture of identical sources, we make use of a first-order Markov chain $A^+(x_i, x_j)$ that can be estimated directly from the symbol sequence $\boldsymbol{x}$. The Markov chain $A^+(x_i, x_j)$ captures the transition probabilities in the observed symbol sequence $\boldsymbol{x}$ as if this sequence had been generated by a single Markov model. The relationship between $A^+(x_i, x_j)$ and $A(x_i, x_j)$ can be obtained by considering that

$$\begin{aligned} P(x_n|x_{n-1}) &= \sum_k \pi_k P(x_n|x_{n-1}, s_n = k) \\ &= \sum_k \sum_l \pi_k \pi_l P(x_n|x_{n-1}, s_n = k, s_{n-1} = l) \\ &= \sum_k \pi_k^2 A(x_{n-1}, x_n) + \sum_k \sum_{l \neq k} \pi_k \pi_l e(x_n) \quad (1) \end{aligned}$$

where π represents the source selection probability and $e(x_n)$ is the steady-state emission probability of x_n. Eq. (1) can be further simplified if we assume that all source probabilities are equal, i.e. $\pi_k = \pi = 1/K$ where K is the number of sources. In that case,

$$A^+(x_{n-1}, x_n) = \pi A(x_{n-1}, x_n) + (1 - \pi) e(x_n) \quad (2)$$

Estimating the discrete Markov model $A^+(x_i, x_j)$ from an example sequence equates to estimating the probabilities p_{ij} from the number of state transitions $i \to j$ observed in sequence $\boldsymbol{x}$. The maximum-likelihood estimate can be shown to be $\hat{p}_{ij} = n_{ij}/n_i$, where n_{ij} is the number of transitions $i \to j$ observed in $\boldsymbol{x}$ and $n_i = \sum_j n_{ij}$ [14]. For a more robust estimate, here we use a Bayesian approach where p_{ij} are estimated as the posterior expectation of P_{ij}. Using a Hyper-Dirichlet prior, this estimate can be shown to be [15]

$$\hat{p}_{ij} = \frac{\alpha_{ij} + n_{ij}}{\alpha_i + n_i} \quad (3)$$

where α_{ij} are hyper-parameters of the prior, selected to represent a prior belief about the distribution. Note that from $A^+(x_i, x_j) = \hat{p}_{ij}$ one can obtain the steady-state distribution for each symbol as $e^+(x_i) = \hat{p}_i$.

If the number of sources K is known and the sources are identical, then it is apparent that $\pi = \frac{1}{K}$ and $e^+(x_i) = e(x_i)$, i.e. the marginal probability of each symbol is the same if we draw the symbols from a single source or from a set of K sources that are all identical to the single one. Eq. (1) can then be reworked to yield,

$$A(x_i, x_j) \quad = \quad K \cdot A^+(x_i, x_j) - (K - 1) \cdot e^+(x_j) \quad (4)$$

Thus, the parameters $A(x_i, x_j)$ of the model can be easily determined from simple observed frequencies when K is known. Due to imprecise estimates of $\hat{p}_{ij}$ and $\hat{p}_j$ from finite sequences, in practice there is the possibility that some elements of $A(x_i, x_j)$ become negative. This can be handled by simply enforcing a minimum value of zero and normalizing:

$$\overline{A}(x_i, x_j) = \begin{cases} \frac{A(x_i, x_j)}{\sum_{j:A(x_i,x_j)>0} A(x_i, x_j)} & A(x_i, x_j) > 0 \\ 0 & A(x_i, x_j) \leq 0 \end{cases} \quad (5)$$

We then use $\overline{A}(x_i, x_j)$ as the estimate for the transition probabilities of the generating Markov models.

B. Estimating the number of sources

Assuming that an example sequence was generated by a mixture of identical Markov models, we can use a straightforward approach to, from this sequence, estimate the most likely number of sources K when the actual number of sources is unknown. We write the probability distribution over K as

$$P(K|\boldsymbol{x}) = \frac{P(K) \cdot P(\boldsymbol{x}|K)}{P(\boldsymbol{x})}$$

meaning that we can find the value of K for which this expression is maximized through

$$\arg\max_{K} = P(K)P(\boldsymbol{x}|K)$$

Using Eq. (5), it is possible to show that $P(\boldsymbol{x}|K)$ can be approximated by

$$\begin{aligned} P(\boldsymbol{x}|K) = \prod_{n} \Bigg[& e(x_n) \cdot \left(1 - \frac{1}{K}\right)^{n-1} + \\ & + \sum_{j} A(x_j, x_n) \cdot \frac{1}{K} \cdot \left(1 - \frac{1}{K}\right)^{n-j-1} \Bigg] \end{aligned} \quad (6)$$

where the sum over j goes over all previous symbols before x_n. Note that the terms of this sum tend to zero as the difference between n and j increases so we can reliably ignore terms with large j to keep the computational complexity low.

Here we will assume the prior distribution $P(K)$ to be a Poisson distribution truncated at the origin (see [16]),

$$P(K) = \frac{\lambda^K}{(e^{\lambda} - 1)K!}, \quad K = 1, 2, \ldots$$

The expected number of components λ is selected in accordance with the prior belief on the number of components in the generating system.

IV. Experiments

The above estimation algorithms will find applications in problems that can be seen as having concurrent Markov processes, for example in process mining [17] where the goal is to find the process model (usually as a Petri net) that best describes the behavior recorded in an event log.

To measure the distance between the generating and estimated models, we will use the *relative entropy rate* that has also been used to measure the distance between two hidden Markov models [18]. Let $D(P_{\beta}^{(N)} || P_{\gamma}^{(N)})$ be the relative entropy between two probability distributions,

$$\begin{aligned} D(P_{\beta}^{(N)} || P_{\gamma}^{(N)}) = \sum_{x_1, \ldots, x_N} & P_{\beta}(x_1, \ldots, x_N) \cdot \\ & \cdot \log \frac{P_{\beta}(x_1, \ldots, x_N)}{P_{\gamma}(x_1, \ldots, x_N)} \end{aligned} \quad (7)$$

Then the relative entropy rate can be defined as:

$$D(\beta || \gamma) \triangleq \lim_{N \to \infty} \frac{1}{N} \cdot D(P_{\beta}^{(N)} || P_{\gamma}^{(N)}) \quad (8)$$

which in turn can be computed through the conditional relative entropy using,

$$\begin{aligned} D(\beta || \gamma) = \lim_{N \to \infty} \sum_{x_1, \ldots, x_N} & P_{\beta}(x_1, \ldots, x_N) \cdot \\ & \cdot \log \frac{P_{\beta}(x_N | x_1, \ldots, x_{N-1})}{P_{\gamma}(x_N | x_1, \ldots, x_{N-1})} \end{aligned} \quad (9)$$

Intuitively, the relative entropy rate can be interpreted as the average information loss per symbol if we compress data generated by β while assuming it was generated by γ. Here we use a Monte-Carlo approach to approximate Eq. (9) by sampling from $P_{\beta}(x_1, \ldots, x_N)$ and calculating the mean estimate of the logarithm.

A. Estimation results on example models

As a first test case, we investigate the similarity between estimated source models and the true source models. Specifically, we compute the estimated model using a symbol sequences produced from an example source model. The purpose of these experiments is to investigate how the similarity between the true and the estimated models depends on the number of sources K and the length N of the input symbol sequence $\boldsymbol{x}$.

In each experiment, a sequence $\boldsymbol{x}$ of length N was drawn from an example model θ' containing K identical sources and having a transition matrix $A'(x_i, x_j)$. From the symbol sequence $\boldsymbol{x}$, a model θ'' with K identical sources and transition matrix $A''(x_i, x_j)$ was estimated under the assumption that K being known. A sample symbol sequence is shown in Fig. 2.

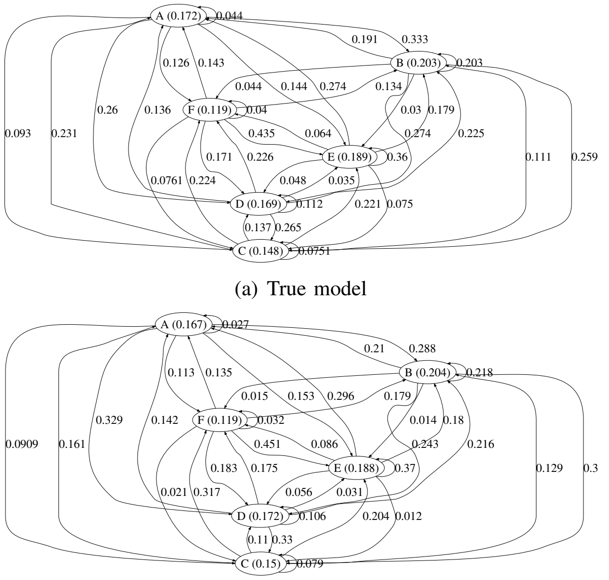
```
DCFCAEEADBEEACDFCFFAFAAFCCDADEBDAADFCFBEEAFDFCAFE
AACADDEBBBCDCFDADEBFFACDFEFFCCCEEBAFCEFBBDCBBEEDF
CECDFDEFCEEEBECBBDCDECACBEFACBEEBCCFBCDDCEFEDAABF
BCDEDACAFEFBAEEBFDEDECDFACAFAEBDCECFCBBEEDE...
```

Fig. 2. Example of an input symbol sequence generated from a model with $K = 5$ sources.

All the experiments were performed based on both deterministic and non-deterministic source models containing 6 different states from A-F. The deterministic models allow only one transition to take place for a given state; that transition has probability 1.0 and all other transitions from the same state have zero probability. In the non-deterministic, or stochastic models, there may be several possible transitions from a given state; the outgoing transition probabilities from a given state add up to 1.0.

The true models used in these experiments – both the deterministic and the stochastic ones – were generated randomly. Fig. 3 shows one of the stochastic models together with its estimated counterpart.

For the purpose of investigating the similarity between the true model θ' and the estimated model θ'', we have used different similarity measures. In general, it is well known that similarity is difficult to measure accurately, since different measures fit different purposes. Therefore, most similarity measures are often used in an ad-hoc manner, depending on the



(a) True model

(b) Estimated model

Fig. 3. Example of a true model and the corresponding estimated model obtained from a generated symbol sequence of length $N = 20000$ symbols when using $K = 5$ sources.

K	N	$D_{\theta',\theta''}$	$\ell(\theta')$	$\ell(\theta'')$	D_{KL}	$D_{\theta'\|\|\theta''}$
5	5000	0.1302	-8681	-8691	7.2043	0.0073
	10000	0.0945	-17330	-17345	3.7571	0.0038
	20000	0.0716	-35019	-35033	1.9728	0.0020
10	5000	0.2509	-8775	-8823	13.8268	0.0140
	10000	0.1583	-17509	-17542	5.6429	0.0056
	20000	0.1336	-35087	-35131	3.3417	0.0034
20	5000	0.3572	-8795	-8871	18.6068	0.0187
	10000	0.3236	-17641	-17759	14.0188	0.0140
	20000	0.2366	-35238	-35382	8.1781	0.0081

TABLE I
SIMILARITY RESULTS BETWEEN ESTIMATED AND TRUE STOCHASTIC MODELS.

K	N	$D_{\theta',\theta''}$	$\ell(\theta')$	$\ell(\theta'')$	D_{KL}	$D_{\theta'\|\|\theta''}$
5	5000	0.1068	-7563	-7661	21.0240	0.0211
	10000	0.0830	-15152	-15285	14.3354	0.0144
	20000	0.0531	-30264	-30435	8.9143	0.0090
10	5000	0.2019	-8325	-8427	22.7431	0.0230
	10000	0.1478	-16580	-16731	15.7120	0.0158
	20000	0.1123	-33171	-33396	11.4560	0.0116
20	5000	0.2969	-8649	-8772	26.2697	0.0265
	10000	0.2373	-17260	-17433	17.4408	0.0177
	20000	0.1991	-34549	-34781	11.8434	0.0120

TABLE II
SIMILARITY RESULTS BETWEEN ESTIMATED AND TRUE DETERMINISTIC MODELS.

application domain. Here, we have used four methods to obtain different similarity measures. The first similarity measure is based on the absolute difference between the models, summed over all state transitions and sources:

$$D(\theta';\theta'') \triangleq \frac{1}{2} \cdot \sum_a \sum_b \mid e'(a) \cdot A'(a,b) - \\ - e''(a) \cdot A''(a,b) \mid \qquad (10)$$

Due to the stochastic constraints that apply to the model parameters $e(a)$ and $A(a,b)$, the value of $D(\theta';\theta'')$ will be at most 1.0. The second similarity measure is simply the log-likelihood $\ell(\theta)$ of the model θ producing the symbol sequence $\boldsymbol{x}$ and is obtained by calculating the probability of observing each new symbol x_n given previous symbols in a sequence. The third similarity measure is based on the Kullback-Leibler divergence D_{KL}, defined in Eq. (7). Finally, the fourth method is based on the reduction in relative entropy $D_{\theta'||\theta''}$ as described in Eq. (9). The results are shown in Tables I and II as the mean of 10 runs.

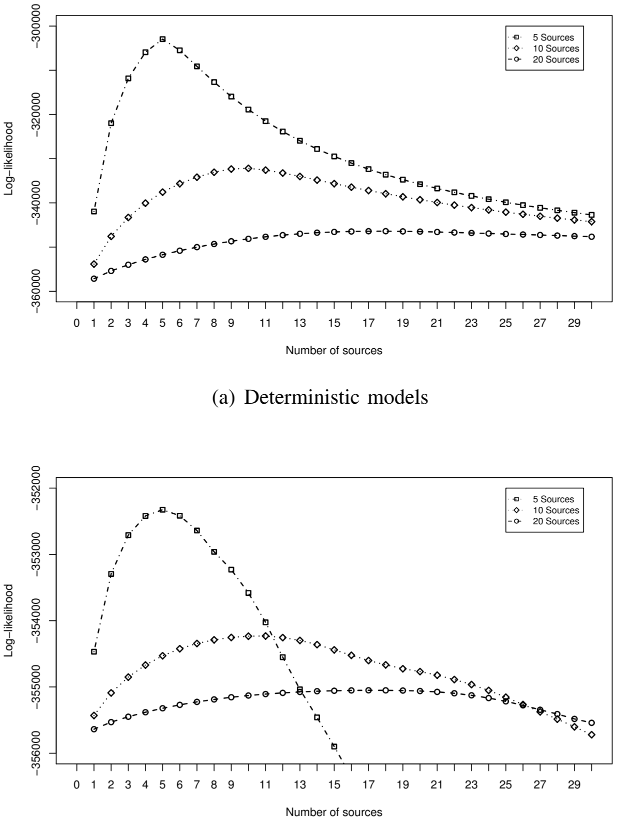
For both stochastic and deterministic source models, we observe that the similarity improves with longer input sequences for all the different types of measures (Tables I and II). Naturally, the estimation accuracy improves with the amount of available data. Further, we observe that the degree of similarity also depends on the number of sources. For example, we see in Table I that for models estimated from a sequence of $N = 5000$ symbols, the difference between true and estimated models increases with K. Finally, we see that the same difference is in general greater for stochastic models rather than for deterministic models. The reason is that in the stochastic case the input symbol sequence must be longer than in the deterministic case in order to achieve the same level of accuracy. In any case, the results suggest that the proposed method is an effective way to rediscover the true model, and that accuracy can be improved by increasing the length of the input symbol sequence.

B. Estimating the number of sources

To illustrate estimation of the number of sources from a symbol sequence, here we will study the model likelihood $P(\boldsymbol{x}|K)$ as a function of the assumed number of sources. The models, both deterministic and stochastic, were generated as in the previous section, and the likelihoods calculated on a sequence of $N = 200000$ symbols, for improved accuracy.

Fig. 4 shows the varying log-likelihoods over different assumed numbers of sources for models with 5, 10, and 20 sources for both deterministic and stochastic models. In the deterministic case, there is a peak in the likelihood at 5 and 10 sources for the models with exactly these numbers of sources. For the model using 20 sources, the likelihood curve is very level but has a peak at 18 sources, somewhat lower than the actual number of sources.

In the stochastic case, estimation is more difficult. The likelihoods peak at 5, 11, and 18 for the models with 5, 10, and 20 sources, respectively. As the likelihood curves are very smooth, the prior would have a definitive impact here, especially if the symbol sequences where shorter. Still, it is possible to achieve a reasonable estimate of the number of sources even for the rather difficult case of the stochastic model.



(a) Deterministic models

(b) Stochastic models

Fig. 4. Model likelihood as a function of the number of assumed sources for synthetic deterministic and stochastic models.

C. Rediscovering the behavior of a technical support process

This case study comes from a medium-sized IT vendor that offers an advanced software platform for rapid application development. The platform is being improved continuously by successive release versions that add new functionality, improve existing features, and correct existing bugs. Besides extensive manual and automated in-house testing, end users have an active role in pointing out desired improvements and problems to be solved. Each request originates a new so-called *issue* that is recorded in the system and handled by the technical support team.

In order to keep track of all issues and to handle them appropriately, the support team stores all the information available in a central database, and it records all changes in state as a solution to the each problem is developed. Fig. 5 illustrates the sequence of states that the handling process of each issue is supposed to go through, according to management guidelines. When a new request is received, it will be recorded in the system as *New*. Then a team member will look at it and check whether it is a duplicate issue, whether it is relevant, what priority level it should be assigned, whether there is enough information for the issue to be handled, whether there are other issues that could be related to the one submitted, etc. In some cases, the issue may end up being *Discarded* or being labeled as *Duplicated*. In most cases, it will follow a mainstream process: the issue is *Assigned* to a specific team member and the state will be changed to *Open* when that team member starts working on it. At this point, it will generally be a matter of time until either a solution or a workaround is found and the issue becomes *Resolved*. An issue is automatically *Closed* when a new product version that includes the bug fix is released.

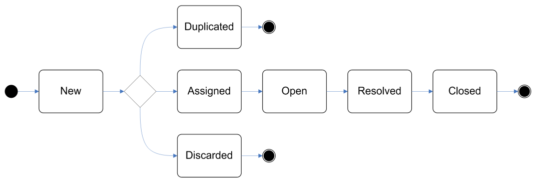


Fig. 5. State changes within the issue handling process.

For the purpose of this case study, we had access to a subset of the event history recorded in the database between September 28, 2006 and September 28, 2007. At its peak, the system reported 284 issues were active simultaneously, for a total of 1211 issues recorded during the observation period of one year. We used all the state changes recorded in the system database as a single, unlabeled sequence of input data. In this symbol sequence the events are given in chronological order, but without any information about the issue (i.e. source) that each event belongs to. The complete input sequence has a length of $N = 3127$ symbols. Also, we assume that we do not know how many sources (i.e. issues) there are. Thus, we estimate all model parameters including the number of sources from the input symbol sequence. In this experiment, the parameter α from Eq. (3) was selected as 300, or roughly one tenth of the length of the observed sequence.

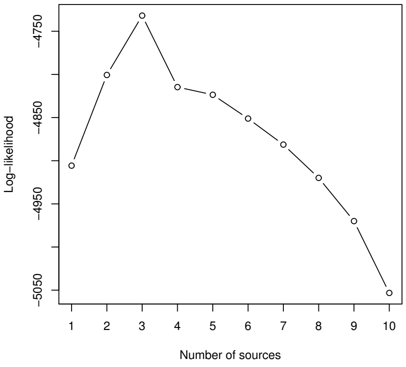


Fig. 6. Model likelihood as a function of the number of assumed sources for the technical support data.

Fig. 6 shows the model likelihood for a varying number of sources. The likelihood clearly peaks for three sources, and then falls as the number of sources increases. The prior over

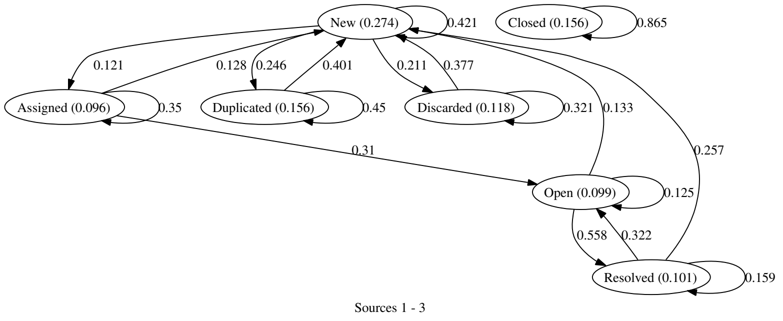


Fig. 7. Estimated model from the technical support data when assuming 3 sources.

the number of components has little effect on this sequence length, thus we estimate the number of concurrent sources to be $K = 3$. Note that the total number of issues is 1211, but 3 sources with repeating behavior seem to be enough to account for the whole symbol sequence.

Fig. 7 shows the resulting estimated model. In the graph, some transitions with very low probability have been removed to improve readability. The model captures the dominant behavior of the issue handling process, with separate paths from *New* to *Duplicated*, *Discarded*, and *Assigned*. After the *Assigned* state, the dominant path is to *Open* and then to *Resolved*. However, this path does not include the *Closed* state as would be expected. This is due to the fact that the behavior associated with the closed state actually does not match the model assumptions at all: in practice, it is only when a large number of issues have been resolved that a new version of the software is released, and at that point all resolved issues are closed simultaneously. Even in this case, the estimated model effectively captures the true behavior of the original process.

V. Conclusions

In this paper we have presented a method to estimate the parameters of randomly interleaved Markov models directly from a symbol sequence under the assumption that all sources are identical. A method to estimate the actual number of sources has also been proposed. Results indicate that when the number of sources or symbols increases, a longer input sequence is required for more reliable estimates.

The algorithm is very fast and scales linearly in the sequence length, meaning that it is suitable for large databases found in practical applications. Additional experiments on both synthetic and real-world data are being carried out to further investigate the performance and accuracy of the proposed method in discovering the hidden logic behind unlabeled symbol sequences.

References

[1] N. Landwehr, "Modeling interleaved hidden processes," in *Proceedings of the 25th International Conference on Machine Learning (ICML-08)*, A. McCallum and S. Roweis, Eds. Omnipress, July 2008, pp. 520–527.

[2] T. Batu, S. Guha, and S. Kannan, "Inferring mixtures of Markov chains," in *Learning Theory*, ser. Lecture Notes on Computer Science. Springer, 2004, vol. 3120, pp. 186–199.

[3] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[4] S. Fine, Y. Singer, and N. Tishby, "The hierarchical hidden Markov model: Analysis and applications," *Machine Learning*, vol. 32, no. 1, pp. 41–62, 1998.

[5] Z. Gharamani and M. Jordan, "Factorial hidden Markov models," *Machine Learning*, vol. 29, pp. 245–275, 1997.

[6] I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White, "Model-based clustering and visualization of navigation patterns on a web site," *Data Mining and Knowledge Discovery*, vol. 7, no. 4, pp. 399–424, October 2003.

[7] E. Manavoglu, D. Pavlov, and C. L. Giles, "Probabilistic user behavior models," in *ICDM '03: Proceedings of the Third IEEE International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2003, p. 203.

[8] S. Frühwirth-Schnatter and C. Pamminger, "Bayesian clustering of categorical time series using finite mixtures of Markov chain models," Department for Applied Statistics, Johannes Kepler University Linz, IFAS Research Paper Series 2007-30, October 2007.

[9] D. Fougere and T. Kamionka, "Bayesian inference for the mover-stayer model in continuous time with an application to labour market transition data," *Journal of Applied Econometrics*, vol. 18, no. 6, pp. 697–723, 2003.

[10] H. Frydman and A. Kadam, "Estimation in the continuous time mover-stayer model with an application to bond ratings migration," *Applied Stochastic Models in Business and Industry*, vol. 20, no. 2, pp. 155–170, 2004.

[11] H. Chen, S. Duffy, and L. Tabar, "A mover-stayer mixture of Markov chain models for the assessment of dedifferentiation and tumour progression in breast cancer," *Journal of Applied Statistics*, vol. 24, no. 3, pp. 265–278, June 1997.

[12] C. Rossi and G. Schinaia, "The mover-stayer model for the HIV/AIDS epidemic in action," *Interfaces*, vol. 28, no. 3, pp. 127–143, March 1998.

[13] C. Fuchs and J. Greenhouse, "The EM algorithm for maximum likelihood estimation in the mover-stayer model," *Biometrics*, vol. 44, no. 2, pp. 605–613, June 1988.

[14] Y. M. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA, 1975.

[15] M. Ramoni and P. Sebastiani, "Bayesian methods," in *Intelligent Data Analysis. An introduction*, M. Berthold and D. J. Hand, Eds. Springer, New York, NY, 1999, pp. 129–166.

[16] R. S and P. J. Green, "On Bayesian analysis of of mixtures with an unknown number of components (with discussion)," *Journal of the Royal Statistical Society B*, vol. 59, pp. 731–792, 1997.

[17] W. van der Aalst, B. van Dongen, J. Herbst, L. Maruster, G. Schimm, and A. Weijters, "Workflow mining: A survey of issues and approaches," *Data Knowledge and Engineering*, vol. 47, no. 2, pp. 237–267, 2003.

[18] B. H. Juang and L. R. Rabiner, "A probabilistic distance measure for HMMs," *AT&T Technical Journal*, vol. 64, no. 2, pp. 391–408, 1985.