Data Association for Topic Intensity Tracking

Andreas Krause **Jure Leskovec** **Carlos Guestrin**

April 2006
CMU-ML-06-100

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

We present a unified model of what was traditionally viewed as two separate tasks: data association and intensity tracking of multiple topics over time. In the data association part, the task is to assign a topic (a class) to each data point, and the intensity tracking part models the bursts and changes in intensities of topics over time.

Our approach to this problem combines an extension of Factorial Hidden Markov models for topic intensity tracking with exponential order statistics for implicit data association. Our approach is general in the sense that it can be combined with a variety of learning techniques; we demonstrate this flexibility by applying it in a supervised, an unsupervised and a semi-supervised (active learning) setting. Experiments on text and email datasets show that the interplay of classification and topic intensity tracking improves the accuracy of both classification and intensity tracking. Even a little noise in topic assignments can mislead the traditional algorithms. However, our approach detects correct topic intensities even with 30% topic noise.

Keywords: Text classification; Dynamic Bayesian Networks; Data Association; Order statistics

1 Introduction

When following a news event, the content and the temporal information are both important factors in understanding the evolution and the dynamics of the news topic over time. When recognizing human activity, the observed person often performs a variety of tasks *in parallel*, each with a different *intensity*, and this intensity *changes over time*. Both examples have in common a notion of classification: e.g., classifying documents into topics, and actions into activities. Another common point is the temporal aspect: the intensity of each topic or activity changes over time.

In a stream of incoming email for example, we want to associate each email with a topic, and then model bursts and changes in the frequency of emails of each topic. A simple approach to this problem would be to first consider associating each email with a topic using some supervised, semi-supervised or unsupervised (clustering) method; thus segmenting the joint stream into a stream for *each* topic. Then, using only data from each individual topic, we could identify bursts and changes in topic activity over time. In this traditional view (Kleinberg, 2003), the data association (topic segmentation) problem and the burst detection (intensity estimation) problem are viewed as two distinct tasks. However, this separation seems unnatural and introduces additional bias to the model. We combine the tasks of data association and intensity tracking into a single model, where we allow the temporal information to influence classification. The intuition is that by using temporal information the classification would improve, and by improved classification the topic intensity and topic content evolution tracking also benefit.

Our approach combines an extension of Factorial Hidden Markov models (Ghahramani & Jordan, 1995) for topic intensity tracking with exponential order statistics for implicit data association. Additionally, we demonstrate the use of a switching Kalman Filter to track content evolution of the topic over time. Our approach is general in the sense that it can be combined with a variety of learning techniques; we demonstrate this flexibility by applying it in supervised, unsupervised and semi-supervised (active learning) settings. Experimental results show that the interplay of classification and topic intensity tracking improves accuracy of both classification and intensity tracking. More specifically, our contributions are:

- A suite of models, EDA–IT, IDA–IT and IDA–ITT, for simultaneous reasoning about topic labels and topic intensities, and extensions to topic drift tracking.
- A modeling trick which uses exponential order statistics to achieve implicit data association. This idea allows us to make an intractable data association problem tractable for exact inference, and is of independent interest.
- The extensive empirical evaluation in the supervised and unsupervised setting on synthetic as well as two real world datasets.

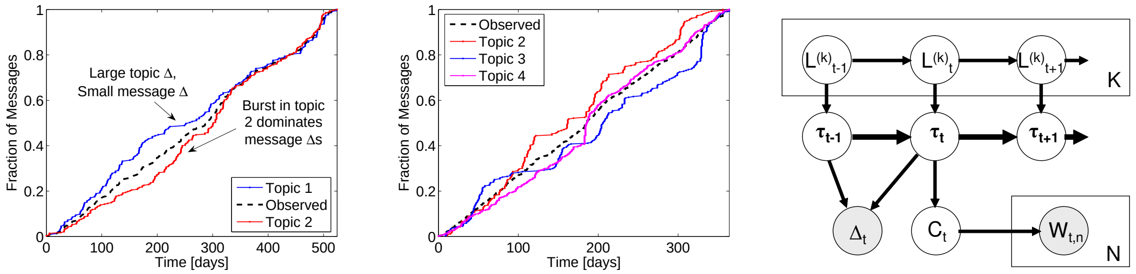
In the following sections we will use email topic detection and tracking as our running example. We also use the terms topic and class as synonyms. Also note that our approach is not limited to the text domain. All our methods are general in a sense that they can be applied to any problem with simultaneous classification and class intensity tracking (e.g., activity recognition).

2 Classification and intensity tracking in the static case

Traditionally, classification refers to the task of assigning a class label c to an unlabeled example x, given a set of training examples x_i and corresponding classes c_i. Classification can be performed by calculating the probability distribution over the class assignments, $P(c|x)$, using Bayes' rule, $P(c|x) \propto P(c)P(x|c)$, where the class prior $P(c)$ and conditional probability of the data $P(x|c)$ are estimated from the training set.

Work in the areas of clustering, topic detection and tracking, e.g., (Allan et al., 1998; Yang et al., 2000), and text mining, e.g., (Swan & Allan, 2000; Blei et al., 2003), has explored techniques for identifying topics in document streams using a combination of content analysis and time-series modeling. Most of these techniques are guided by the intuition that the appearance of a topic in a document stream is signaled by a *burst*, a sharp increase of intensity of document arrivals. For example, in the problem of classifying emails into topics, the focus of attention might change from one topic to another and hence taking into account the topic intensity should help us in the classification task.

To define the notion of *intensity*, consider a task where we are given a sequence of n email messages, $x_1, \ldots, x_n$, and are asked to assign a topic c to each email. We also observe the message arrival times $t_1, \ldots, t_n$. The *intensity* λ_c of a topic c is defined as the *rate* at which documents of that topic appear, or



(a) Enron data

(b) Reuters data

(c) EDA-IT

Figure 1: Topic deltas and observed message deltas for email (a) and news data (b). Note how observed deltas are dominated by bursts in a single topic. Observed deltas in (b) look almost uniform, despite strong bursts in topic intensities. Explicit (but intractable) data association model (c) capturing the intensity-driven generative process. We observe t-th message from the distribution over N words and Δ_t, the elapsed time from the last received email. We have K topics, each with intensity $L_t^{(k)}$ at time t. C_t is the topic indicator, τ_t stores the time of last email of each topic.

equivalently as the inverse expected interarrival time $\mathbb{E}[\Delta_c]^{-1}$ of the topic c, where $\Delta_{c,i} = t_{c,i} - t_{c,i-1}$ is the time difference between two subsequent emails from the same topic c. A natural model of interarrival times is the exponential distribution (Kleinberg, 2003), i.e., $\Delta \sim \text{Exp}(\lambda)$, with density $p(\Delta \mid \lambda) = \lambda \exp(-\lambda\Delta)$.

Let us first consider the case of a single topic. A naïve solution to estimating intensity dynamics would be to compute average intensities over fixed time windows. Since the exponential distribution has very high variance, this procedure is likely not to be very robust. Furthermore, it is not easy to select the appropriate length for the time window, since, depending on the topic intensity, the same time window will contain very different numbers of messages. Also, from the perspective of identifying bursts in the data, a set of discrete levels of intensity is preferable (Aizen et al., 2004). To overcome these problems, Kleinberg (2003) proposed a *weighted automaton model (WAM)*, an infinite-state automaton, where each state corresponds to a particular discrete level of intensity. For each email, a transition is made in the automaton, whereby changes in intensities are penalized. This can be interpreted as a Hidden Markov Model, where the search for the most likely parameters of the exponentially distributed topic deltas $\Delta_{c,i}$ reduces to the Viterbi algorithm.

Since the WAM model operates on a single topic only, hard assignments of messages to topics have to be made in advance. Although classification can be done using methods as described in (Blei & Lafferty, 2005; Segal & Kephart, 1999), these hard assignments imply that topic detection and identification of bursts are separated. However, our intuition is that temporal information should help us assign the right topic and that the topic of an email will influence topic intensity. For example, if we are working on a topic with a very high intensity and the next email arrives at the right moment, then this will influence our belief about the email's topic. On the other hand, if an email arrives late and we are very sure about its topic, we will have to revise our belief about the intensity of the topic.

In the following sections, we propose a suite of models which simultaneously reason about topic labels and topic intensities. In Section 6 we show how a little class topic assignment noise can confuse WAM, while our model still identifies the true topic intensity level.

3 Classification and intensity tracking in the dynamic case

Given a stream of data points (we can think of them as emails) on K topics (classes) together with their arrival times, (x_1, t_1), (x_2, t_2), (x_3, t_3), $\ldots$, we want to simultaneously classify the emails into topics and detect bursts in the intensity of each of the topics.

We have a *data association* problem: We observe the *message deltas* $\Delta_i = t_i - t_{i-1}$, the time between arrivals of consecutive emails. One first needs to associate each email with a correct topic to find the *topic deltas*, the time between messages of the *same* topic. Given the topic deltas one can then determine the *topic intensity*.

For example, Figure 1(a) shows arrival times for email data and indicates importance of the data association part. Each dot represents an email message and we plot the message number vs. the time of a message. Vertical parts of the plot correspond to bursts of activity. Horizontal parts correspond to low activity (long time between consecutive emails). Not knowing the true topics, we only observe the black dotted curve in the middle and we need to associate each email with the correct topic (curves above and below the middle one). Notice how bursts in activity of one topic dominate the observed deltas (middle dotted curve).

A naïve approach to solving the data association problem described above is to explicitly keep track of when we last saw a message from a given topic. Figure 1(c) presents a Dynamic Bayesian Network (DBN) for modeling such a process. Each topic k is associated with an *intensity process* $L_t^{(k)}$, which is a Markov chain modeling the change of topic intensity over time. The discrete states $L_t^{(k)} = l$ are associated with a parameter $\lambda(l)$ of an exponential distribution, modeling the message interarrival times for topic k. We model the topic transition probabilities as $P(L_{t+1}^{(k)} = l \mid L_t^{(k)} = l) = 1 - \theta$, and $P(|L_{t+1}^{(k)} - l| = 1 \mid L_t^{(k)} = l) = \theta/2$, properly accounting for boundary cases. So we allow intensity to increase or decrease with probability θ, which is a parameter of the model.

We can *explicitly* model $\tau_t^{(k)}$, the time at which we last saw an email from topic k, as a vector $\tau_{\mathbf{t}}$. At each time index t, the topic c_t of the t-th message is $c_t = \text{argmax}_k\, \tau_t^{(k)}$, i.e., the last message on topic c_t happened at the current time. The transition from $\tau_{\mathbf{t}}$ to the next time step $\tau_{\mathbf{t+1}}$ is a follows: for each topic k, a new arrival time $\tau'^{(k)}_{t+1}$ is generated by incrementing $\tau_t^{(k)}$ by the *topic delta*, which is exponentially distributed with parameter $\lambda(L_t^{(k)})$. Now, the smallest of τ'_{t+1} determines the email arrival time and the index determines the topic, $c_{t+1} = \text{argmin}_k\, \tau'^{(k)}_{t+1}$. For the topic c_{t+1} of this new email, we update the last topic access time $\tau_{t+1}^{(c_{t+1})} = {\tau'}_{t+1}^{(c_{t+1})}$. The remaining $\tau_{t+1}^{(k)}$ for $k \neq c_{t+1}$ are unchanged, and remain identical to $\tau_t^{(k)}$.

In our problem the model observes *message delta*, $\Delta_t = \max_k \tau_t^{(k)} - \max_k \tau_{t-1}^{(k)}$, which is the time between the current message and the previous one. We also observe a representation w_t of the message, e.g., a bag of words representation. Unfortunately, the inference in this model is intractable – the state space grows as T^K with the number T of documents. Conceptually, the explicit data association model EDA–IT, as sketched in Figure 1(c), represents the generative process, underlying the intensity driven generation of document streams. Instead of investigating heuristics for coping with the intractability of the presented model, we now introduce a simpler model, which elegantly avoids the intractability of explicit data association.

4 Implicit data association models

4.1 IDA-IT: Supervised, implicit data association for intensity tracking

From Section 3 we have that the topic c_{t+1} of next message is the one with minimum ${\tau'}_{t+1}^{(k)} = \tau_t^{(k)} + \Delta_k$. Here $\Delta_k \sim \text{Exp}[\lambda(L_k)]$ is the *topic delta*, i.e., the time between consecutive emails from topic k. So the probability that the next email is from topic k is $P(c_{t+1} = k) = P(\tau_t^{(k)} + \Delta_k \leq \tau_t^{(j)} + \Delta_j \mid \tau_t^{(j)} + \Delta_j \geq r)$, where j ranges over all topics, and $r = \max_k \tau_t^{(k)}$ is the arrival time of email at time index t. So, the probability that the topic of next arriving email is k, is the chance that ${\tau'}_{t+1}^{(k)}$ is the earliest of all "scheduled" arrivals, conditioned on how much time has passed.

The key to making the EDA–IT model tractable is to exploit the *memorylessness* of the exponential distribution to avoid keeping track of the times $\tau_{\mathbf{t}}$ when we have last seen a message on each topic. The memorylessness property states that, if $X \sim \text{Exp}(\lambda)$, then $P(X > T + t \mid X > T) = P(X > t)$. Assuming the intensities of each topic are fixed, it follows that

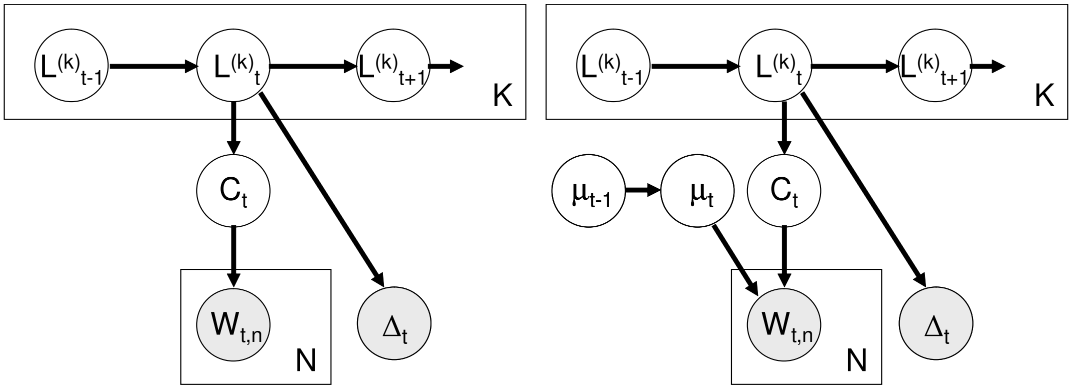
$$(C_t|\mathbf{L_t} = \mathbf{l}) \sim \text{argmin}\{\text{Exp}[\lambda(l_1)], .., \text{Exp}[\lambda(l_k)])\}, \tag{1}$$

$$(\Delta_t|\mathbf{L_t} = \mathbf{l}) \sim \min\{\text{Exp}[\lambda(l_1)], .., \text{Exp}[\lambda(l_k)])\}. \tag{2}$$

Both conditional probability distributions (CPDs) rely on *exponential order statistics*: The observed message delta is the minimum of several exponential distributions (Eq. 2), whereas the selected topic is the corresponding index of the smallest variable (Eq. 1). At first glance, since these CPDs represent complex order statistics, it is not obvious whether they can be represented compactly and evaluated efficiently. The following result (Trivedi, 2002) gives simple closed form expressions for the CPDs 1 and 2:

Proposition 1 *Let* $\lambda_1, \ldots, \lambda_n > 0$ *and* $Z_1 \sim \text{Exp}(\lambda_1)$, $\ldots$, $Z_n \sim \text{Exp}(\lambda_n)$. *Then* $\min\{Z_1, \ldots, Z_n\} \sim \text{Exp}(\sum_j \lambda_j)$ *and* $P(Z_i = \min\{Z_1, \ldots, Z_n\}) = \frac{\lambda_i}{\sum_j \lambda_j}$.

Using these CPDs, we arrive at the model presented in Figure 2(a). We retain the intensity processes $L_t^{(k)}$, but instead of keeping track of $\tau_{\mathbf{t}}^{(\mathbf{k})}$, the time of last email of each topic, and deriving the topic label c_t



(a) IDA–IT

(b) IDA–ITT

Figure 2: Proposed graphical models. (a) Implicit (and tractable) data association and intensity tracking; (b) Implicit data association with intensity and topic tracking.

from it, we use the intensities $\mathbf{L_t}$ directly to model the topic prior. In this model, the association of message deltas (time between consecutive emails) to topic deltas (time between consecutive emails of the *same* topic) is *implicitly* represented. We refer to this model as IDA–IT, *Implicit Data Association for Intensity Tracking.*

The order statistics simplification is an approximation, since in general the topic intensities are not constant during the interval between emails. Our model makes the simplifying assumption that the topic is conditionally independent of the message delta given the topic intensities. However, our experimental results indicate that this approximation is very powerful and performs very well in practice. Moreover, the IDA–IT model now lends itself to exact inference (for a small number of topics). IDA–IT is a simple extension of the Factorial Hidden Markov Model (Ghahramani & Jordan, 1995), for which a large variety of efficient approximate inference methods are readily available. Note that the IDA–IT model is a special case of continuous time models such as continuous time Bayesian Networks (CTBNs) (Nodelman et al., 2003). Unlike our model, CTBNs are in general intractable, and one has to resort to approximate inference (*c.f.*, Ng et al., 2005).

4.2 IDA–ITT: Unsupervised topic and intensity tracking

In a truly dynamic setting, such as a stream of documents, we do not only expect the topic intensities to change over time, but the vocabulary of the topic itself is also likely to change, an effect known as *topic drift.* Next, we present an extension of IDA–IT model that also allows for tracking the evolution of the content of the topics.

Here we use the Switching Kalman Filter to track the time evolution of the words associated with each topic. We represent each topic with its centroid – a center of the documents in the topic. As the topic content changes, the Kalman filter tracks the centroid of the topic over time. Since representing documents in the bag–of–words fashion results in extremely high dimensional spaces, where modeling topic drift becomes difficult, we adopt the commonly used Latent Semantic Indexing (Deerwester et al., 1990) to represent documents as vectors in a low dimensional space.

Using the Gaussian Naïve Bayes model, the observation model for documents becomes $P(W_{t,i} \mid C_t = k) \sim \mathcal{N}(\mu_{i,k}, \sigma^2_{i,k})$, where we represent each topic by its mean $\mu^{(\mathbf{k})}$ and variance $\sigma^{(\mathbf{k})}$. For simplicity of presentation, we will assume that only the topic centers change over time, while variances remain constant. Assuming a normal prior on the mean, and a normal drift, i.e., $\mu^{(\mathbf{k})}_{\mathbf{t+1}} = \mu^{(\mathbf{k})}_{\mathbf{t}} + \nu$ for $\nu \sim \mathcal{N}(0, \varepsilon^2)$, we can model the topic drift $\mu^{(\mathbf{k})}_{\mathbf{1}}, \ldots, \mu^{(\mathbf{k})}_{\mathbf{T}}$ by plugging a Switching Kalman Filter (SKF) into our IDA–IT model. We call this model Implicit Data Association for Intensity and Topic Tracking (IDA–ITT), presented in Figure 2(b).

The SKF model fits in the following way: The continuous state vector $\mu_t = (\mu_t^{(1)}, \ldots, \mu_t^{(K)})$ describes the prior for the topic means. The linear transition model is simply the identity, i.e., $\mu_{t+1} = \mu_t + \nu$. This means that we expect the prior to stay constant, but allow a small Gaussian drift ν. The observation model is a Gaussian distribution dependent on the topic: $W_t \mid [\mu_t, C_t = c] \sim \mathcal{N}(H_c \cdot \mu_t, \Sigma_c)$. Hereby, H_c is a matrix selecting the mean $\mu_t^{(c)}$ from the state vector μ_t. For example, in the case of two classes, and the documents represented as points in $\mathbb{R}^2$, $H_1 = (1,1,0,0)$ and $H_2 = (0,0,1,1)$. We can estimate Σ_c from training data and keep it constant, or associate it with a Wishart prior. In this paper, we select the first option for clarity of presentation.

Unfortunately, we cannot expect to do exact inference anymore, since inference in such hybrid models is intractable (Lerner & Parr, 2001). However, there are very good approximations for inference in Switching Kalman Filters (Lerner, 2002). We will briefly explain our approach to inference in Section 4.5.

4.3 Active Learning for IDA–ITT

We also extended of our model to the semi-supervised, expert-guided classification case, where occasional expert labels for the hidden variables are available, and investigated an active learning method for selecting most informative such labels.

The extension of our models to the semi-supervised case is straight-forward: Since we maintain a full probabilistic model over the intensity processes and the class labels, we can incorporate expert labels by performing probabilistic inference. Since in most applications, expert labels are expensive, it is necessary to determine for which unseen examples we should request the label in order to most effectively improve in the classification task. This procedure is called *active learning.* In our setting the expert will only look at isolated emails, so we can only request labels for the class variables C_t, and not for the intensity variables L_t. I.e., in the email classification example, the expert can query the topic of one of the emails in the stream. The intuition is that if the email is in a burst, the other emails close by will likely be from the same topic. The particular outcome of the observations we make will determine our belief about topic labels of other emails. Hence, we want to make our next request dependent on which labels we already received. Unfortunately, the number of sequences of m labels to consider grows exponentially in m. However, in (Krause & Guestrin, 2005) it was shown that for chain graphical models, such as HMMs, this optimal *conditional plan* to decide which hidden variables to observe, depending on the observations already made, can be efficiently computed.

In order to quantify the uncertainty in the topic, we use the conditional Shannon entropy,

$$H(C_t \mid \mathbf{O}_{1:t}) = -\sum_{c,\mathbf{o}} P(C_t = c, \mathbf{O}_{1:t} = \mathbf{o}) \log_2 P(C_t = c \mid \mathbf{O}_{1:t} = \mathbf{o}),$$

whereby $\mathbf{O}_{1:t}$ denotes the observations made for emails 1 through t. The set of observations $\mathbf{O}_{1:t}$ consists of the message deltas $\Delta_{1:t}$ and document representations $W_{1:t}$, and additionally some observed topic labels $C'_{1:t} = \{C_{i_1}, \ldots, C_{i_m}\}$ which we choose to observe. We assume there is a budget function $B(t)$ which increases monotonically with t, and which specifies how many topic labels can be queried up to time t, i.e., we require $|C'_{1:t}| \leq B(t)$ for all t. The objective we optimize is

$$\min_{C'} \mathbb{E}\left[\sum_{t=1}^{T} H(C_t \mid C'_{1:t}, \Delta_{1:t}, W_{1:t})\right], \tag{3}$$

the expected sum of conditional entropies $H(C_t \mid C'_{1:t}, \Delta_{1:t}, W_{1:t})$, conditional on the observed deltas $\Delta_{1:t}$ and documents $W_{1:t}$, as well as the choice of class labels $C'_{1:t} \subset \{C_1, \ldots, C_t\}$. Since only the class labels C_t, and not the intensities L_t are observed, the conditional independence assumptions made in (Krause & Guestrin, 2005), namely that C_i is conditionally independent of C_k given C_j if $i < j < k$ do not hold here. Hence a direct application of their dynamic programming algorithm is not guaranteed to provide an optimal solution anymore. Here, as an approximation, we simply use their algorithm as if this conditional independence would hold. Furthermore, in order to avoid having to sum over exponentially many possible observations $O_{t+1:T}$, we sample a fixed number of observation trajectories, and compute (3) by approximating the expectation using these sample trajectories. In our experiments, we found that inspite of these two approximations, appropriate observation selections can still be obtained.

4.4 Generalizations

Our approach is general, in at least three ways. Firstly, as argued in Section 1, the application is not limited to document streams. Another possible application of our models is fault diagnosis in a system of machines with different failure rates, or activity recognition, where the observed person is working on several tasks in parallel with dynamic intensities. Secondly, our models fit well in the supervised, unsupervised and semi-supervised case as demonstrated in the paper. Lastly, instead of using a Naïve Bayes classifier as done here, any other generative model for classification can be "plugged" into our model, such as TAN trees (Friedman et al., 1997) or more complex graphical models. Instead of using Latent Semantic Indexing to represent documents, it is possible to use topic mixture proportions computed using Latent Dirichlet Allocation (LDA) (Blei et al., 2003) or some other method. In the LDA example, one can either apply the SKF directly to the numerical topic mixture proportions, or track the mixture proportions using the Dirichlet distribution (which makes inference more difficult).

Most generally, our model can be considered as a principled way of adapting class priors according to class frequencies changing over time. Instead of assuming that the transition probabilities θ stay constant between any two subsequent events, a possible extension is to let them depend on the actual observed message deltas, by modeling $L_t^{(k)}$ as continuous-time Markov processes. We experimented with this extension, but did not observe significant difference in the behavior, since in our data sets the actual observed deltas were rather uniform (Figure 1(b)). Similarly, the Gaussian topic drift ν in the IDA–ITT model can be made dependent on the observed message delta, allowing larger drifts when the interval between messages is longer.

4.5 Scalability and implementation details

For a small number of topics, exact inference in the IDA–IT model is feasible. The variables $L_t^{(k)}$ and C_t are discrete, and the continuous variables are all observed. Hence, the standard forward-backward and Viterbi algorithm for Hidden Markov Models can be used for inference. Unfortunately, even though the intensity processes $L_t^{(k)}$ are all marginally independent, they become fully connected upon observing the documents and the arrival times, and the tree-width of the model increases linearly – the complexity of exact inference increases exponentially – in the number of topics. Exact inference has complexity $\mathcal{O}(TK^2|L|^{2K})$, where L is the set of intensity levels, and K, T are the number of topics and documents, respectively. However, there are several algorithms available for approximate inference in such Factorial Hidden Markov Models (Ghahramani & Jordan, 1995). We implemented an approach based on particle filtering, and fully-factorized mean field variational inference. In Section 6, we present results of our comparison of these methods with the exact inference.

For the topic tracking model IDA–ITT, our implementation was based on the algorithm for inference in SKFs proposed by Lerner (2002). The algorithm maintains, at each time step, a belief state over possible locations of the class centers. This belief state is represented by a mixture of Gaussians, where each Gaussian is associated with a class and a vector of intensities:

$$P(x, C_t, L_t \mid \Delta_{1:t}, W_{1:t}) \approx \sum_i s_{t,i} \mathcal{N}(x_t; \mu_i, \Sigma_i),$$

where the mixture weights $s_{t,i}$ sum up to 1 for each time step t. The algorithm then repeatedly performs the following time and measurement update steps: During the time update, the distribution $P(x_{t+1}, C_{t+1}, L_{t+1} \mid \Delta_{1:t}, W_{1:t})$ is computed by multiplying in the transition model and marginalizing out the hidden variables from time step t. This is done exactly with respect to the approximation from the previous time step, i.e., the number of mixture components increases. In the measurement update step, the distribution $P(x_{t+1}, C_{t+1}, L_{t+1} \mid \Delta_{1:t+1}, W_{1:t+1})$ is computed by conditioning each mixture component on the observations Δ_{t+1} and w_{t+1} and re-normalization of the weights s_i. Finally, the mixture is collapsed into a mixture with fewer components using a strategy proposed by Lerner (2002). In our implementation, we keep the four components with the largest weight from each class and each intensity.

5 Experimental setup

Synthetic datasets. First, we evaluate our models on two synthetic datasets. The first dataset (S1) was designed to test whether implicit data association recovers true topic intensity levels. For each of the two topics, we generated a sequence of 300 observations, with exponentially distributed time differences. Every hundred samples, we changed the topic intensity, in the sequence $[\frac{1}{4}, \frac{1}{128}, \frac{1}{32}]$ for topic 1 and $[\frac{1}{128}, \frac{1}{32}, \frac{1}{4}]$ for topic 2. The observed feature W_t is a noisy copy of the topic variable C_t, taking the probability 0.9 for correct topic to introduce additional classification uncertainty.

The second dataset (S2) tests the resilience towards noise in the assignments of messages to topics. Observations in the dataset are uniformly spaced four hours apart, so the observed message deltas are completely uninformative. Every fourth email is from topic 2, the remaining emails are from topic 1. So, the true intensity of topic 1 is $\frac{1}{5}$, and for topic 2 it is $\frac{1}{16}$. We again observe a noisy copy of the true topic label. However, for 30% of the observations from topic 2, the evidence points to the wrong topic – we assign the probability of the correct topic to 0.49; thus hard-assignment of topics will misclassify 30% of messages from topic 1.

Enron email corpus. The Enron dataset contains 517,431 emails from 151 Enron employees. We selected all 554 email messages from *tech–memos* and *universities* folders of employee Kaminski, treating each folder as a separate topic. The email data spans from December 1999 to May 2001.

Reuters document corpus Volume 1 contains 810,000 English language news articles, spanning a year starting from August 1996. We selected 2,303 documents from four topics (*wholesale prices, environment issues, fashion*, and *obituaries*). The number of documents per topic varies between 259 and 938. For each document we also know the time of publication.

Document representation and training. In both real datasets we removed stop-words and words with document frequency of less than 5. We also applied Latent Semantic Indexing (Deerwester et al., 1990) retaining 8 latent dimensions, with components determined on the training data. We decreased the dimensionality of the data to increase interpretability of the results, avoid over-confidence of Naïve Bayes and decrease the number of estimated parameters. In all experiments, we used the first 25% of data for training and the rest for testing. In the Enron data set, this amounts to the first six months of data for training and in Reuters only for a month and a half. Since the documents are not evenly distributed over time and some topics have high (low) intensity at the start of the datasets, the learned class (topic) priors may be different from the true class priors, an issue not addressed by traditional methods.

6 Experimental results

6.1 Topic intensity tracking

Experiments on synthetic datasets. First, we analyze the recovery of the intensity changes on the synthetic dataset. We chose the intensity levels $\frac{1}{4}$, $\frac{1}{16}$, $\frac{1}{32}$, $\frac{1}{64}$, $\frac{1}{128}$ and $\frac{1}{256}$. So the three "correct" intensity levels are available, as well as three "wrong" levels. We set the intensity transition probability to 0.2.

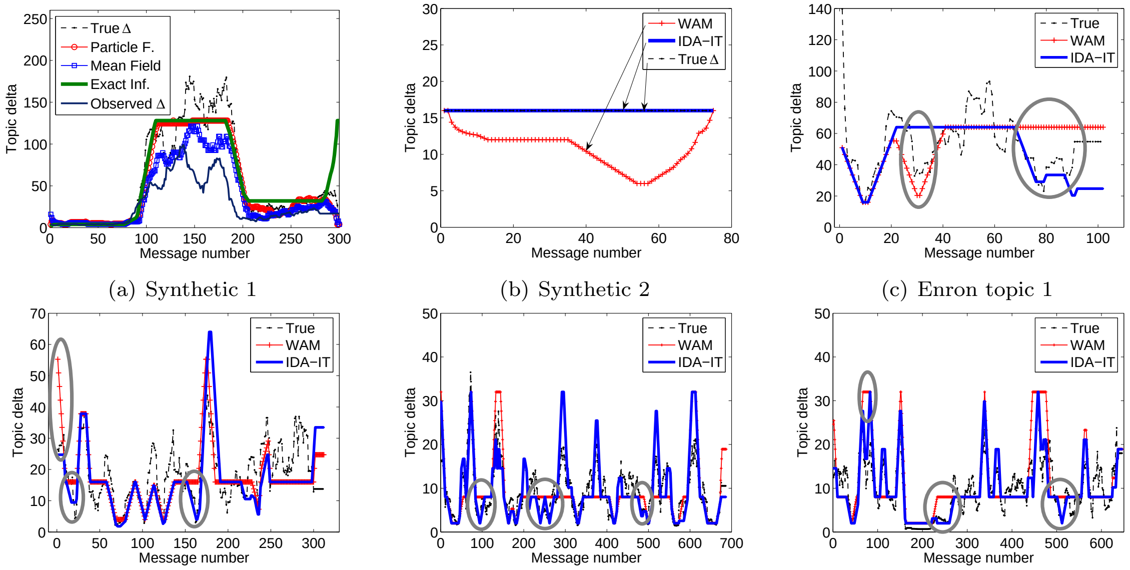
Figure 3(a) presents the results. The x-axis presents documents ordered in time of arrival and on the y-axis we plot the *inverse intensity (average topic delta)*, i.e., time between two consecutive emails from the *same* topic. The dashed lines correspond to the *ground truth* (topic deltas), which are not observed by the algorithms. Notice that the exact inference successfully recovers the true intensities, in spite of the high variance of the exponential distribution. Also observe that the Viterbi decoding successfully avoids simply matching the observed message deltas. This indicates that IDA–IT model succeeds in the data association task. Also, at the end of the sequence, where no messages of topic 2 are observed, the intensity of the low frequency topic is estimated as low, which means we successfully incorporate the "negative evidence".

Figure 3(a) also compares the performance of different inference algorithms to estimate the latent intensities. Both the exact inference and the particle filter recover the true parameters very well. The variational approximation still captures the qualitative behavior, but does not provide as good results as the other methods. This shows that approximate inference can be used for scaling up the model to larger datasets.

Next, we analyze the intensity tracking in presence of classification noise using the synthetic dataset 2, where 30% of examples are misclassified. We compare against the Weighted Automaton Model (WAM) on hard-assigned labels. We chose the intensity levels $\frac{1}{5}$ (correct for topic 1), $\frac{1}{6}$, $\frac{1}{12}$ (both wrong, indicate misclassification) and $\frac{1}{16}$ (correct for topic 2). The intensity transition probability is set to 0.1.

IDA–IT recovers the true underlying rate of both topics. Figure 3(b) shows this for the low intensity topic 2. Estimating the rates after hard-assigning labels drastically decreases performance. Furthermore, all 30% examples misclassified by the Naïve Bayes are correctly classified during our inference. This indicates synergetic effects between intensity estimation and topic identification. It also shows that IDA–IT does true data association of topic deltas, even with completely uninformative message deltas.

Experiments on Enron and Reuters. We compare IDA–IT and the traditional WAM model on Enron and Reuters datasets. Figures 1(a) and 1(b) show the observed data for Enron and Reuters. We plot the



(a) Synthetic 1 (b) Synthetic 2 (c) Enron topic 1

(d) Enron topic 2 (e) Reuters topic 1 (f) Reuters topic 4

Figure 3: (a) True and recovered topic intensities (topic deltas) using various inference techniques with IDA–IT on synthetic dataset 1. (b) Classification noise confuses traditional approach of separate classification and topic intensity tracking. By coupling classification and intensity tracking, IDA–IT recovers true topic intensities. (c)-(f): Comparison of IDA–IT and WAM on Enron and Reuters datasets. We plot intensity level vs. message number. Dashed line presents true intensity and solid lines present recovered intensity level. We circled the areas where WAM model significantly deviates from the truth. Only in one case (first ellipse in (c)) does WAM perform better.

message number versus the time of the message. Each dot represents a message. Vertical parts correspond to bursts of activity and horizontal to low activity (long time between consecutive messages). The algorithm observes the dashed curve in the middle and needs to associate each message with the correct topic (curves above and below the middle one). Notice how bursts in activity of one topic dominate the observations. Notice how in the Reuters data set (Figure 1(b)) the observed message deltas are almost uniform, but the individual topics exhibit strong bursts of activity (sharp vertical jumps on the plot).

Figure 3 shows the results on intensity tracking. Figures 3(c) and 3(d) compare our IDA–IT with the traditional approach, where each message is first assigned a topic and then WAM is run separately on each topic. We circled the spots where the WAM model gets confused due to misclassifications and determines the wrong intensity level. On the contrary, IDA–IT can compensate for classification noise and more accurately recover true intensity levels.

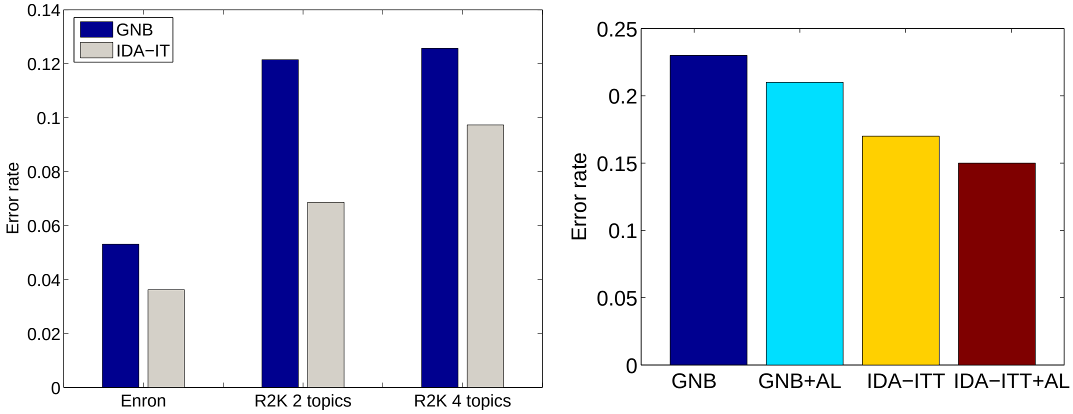
Similarly, Figures 3(e) and 3(f) show the results for 2 out of 4 topics from the Reuters data set. Notice how topic 1 interchanges the low and high activity and how using hard classification with WAM model misses several transitions between intensities. In a data-mining application aiming at the detection of bursts, these lapses would be highly problematic.

6.2 Improved classification

In the previous section, we showed how coupling classification and intensity tracking better models the intensity than if classification and tracking are done separately. Next, we evaluate how classification accuracy is influenced by combining it with intensity tracking.

Figure 4 compares the overall classification error of the baseline, the Gaussian Naïve Bayes classifier, with the proposed IDA–IT model. We ran 3 experiments: Enron emails, topics 1 and 2 from Reuters and all 4 Reuters topics. We used same preprocessing of the data as in the other experiments (see Section 5). For Enron we determined a set of intensity levels 1, $\frac{1}{4}$, $\frac{1}{16}$, $\frac{1}{64}$ and the transition probability of 0.1 using cross-validation. The error rate of Gaussian Naïve Bayes (GNB) is 0.053, IDA–IT scores 0.036, which is a 32% relative decrease of error.

We ran two experiments with Reuters. For both experiments we used intensity levels $\frac{1}{2}$, $\frac{1}{8}$, $\frac{1}{32}$ and transition probability 0.2. In first experiment, we used only topics 1 and 2. Classification error of GNB is



(a) Reduction in error for Enron and Reuters

(b) Active Learning

Figure 4: (a) Error reduction of IDA–IT and Gaussian Naïve Bayes classifier (GNB). Lower is better. Notice how IDA–IT benefits from temporal evidence for classification. (b) Both GNB and IDA–ITT benefit from Active Learning.

0.121 and error of IDA–IT is 0.068, which means 45% relative decrease in classification error. The second experiment uses 4 topics, so the overall performance of both classifiers is lower, but we still get 22% relative improvement in classification.

Note that our models do not have an explicit class prior but model it through topic intensity. This has the effect that the topic which is currently at high activity also has higher prior topic probability. Therefore the precision of the bursty topic increases at the cost of reduced recall for of classes with lower intensity. This usually leads to overall improvement of classification accuracy, but there are cases where improvement is marginal or even decreases due to the lower recall on low intensity topics.

6.3 Topic tracking in unsupervised case

Next, we present the application of implicit data association and intensity tracking model to the unsupervised setting, where we are using the Switching Kalman Filter as introduced in IDA–ITT (section 4.2). For this experiment we chose two Reuters topics, *wholesale prices* and *environment issues*. Using LSI, we reduce the dimensionality of data to two dimensions. We then represent each document as a point in this two-dimensional space and use IDA–ITT to track the evolution of content and intensity of the topics.

Exploring the most important words from the cluster centroid of topic *wholesale prices*, measured by magnitude of LSI coefficients, we see that words *economist, price, bank, index, industry, percent* are important throughout the time. However, at the beginning and the end of the dataset, important words are also *bureau, indicator, national, office, period, report*. Then for few weeks in December and early January the topic drifts towards *expected, higher, impact, market, strong*, which are terms used when last year's trends are analyzed and estimates for next year are announced.

6.4 Active learning

Last set of experiments compares the performance of Naïve Bayes and IDA–ITT, each with and without active learning, using 5% of expert labels. We took topics 1 and 2 from Reuters, and trained only on the first 4% of the data. At each time step, the active learning applies the subset selection algorithm from (Krause & Guestrin, 2005) for a finite look-ahead window of 15 documents. If the subset contains the class label at time t, the budget for requesting labels is decreased by one, and an expert label is obtained for the current document. Per time-step, the budget increases by 0.05 observations, so a new label can be requested on average every 20 documents. Figure 4(b) presents the results, which indicate that intensity and topic tracking as well as Active Learning, as described in Section 4.3, can effectively improve the classification results.

7 Conclusion

We presented a general approach to simultaneous classification of a stream of datapoints and identification of bursts in class intensity. Unlike the traditional approach, we simultaneously addresses data association (classification, clustering) and intensity tracking.

We showed how to combine an extension of Factorial Hidden Markov models for topic intensity tracking with exponential order statistics for implicit data association, which allows efficient inference. Additionally, we applied a switching Kalman Filter to track the time evolution of the words associated with each topic.

Our approach is general in the sense that it can be combined with a variety of learning techniques. We demonstrated this flexibility by applying it in a supervised, unsupervised and a semi-supervised (active learning) setting. Extensive evaluation on real and synthetic datasets showed that the interplay of classification and topic intensity tracking improves the accuracy of both classification and intensity tracking.

Acknowledgements. We would like to thank David Blei for helpful discussions. This work was supported by NSF Grants No. CNS-0509383, IIS-0209107, IIS-0205224, INT-0318547, SENSOR-0329549, EF-0331657, IIS-0326322, Pennsylvania Infrastructure Technology Alliance (PITA) and a donation from Hewlett-Packard. Carlos Guestrin was partly supported by an Alfred P. Sloan Fellowship. Jure Leskovec was partly supported by a Microsoft Research Graduate Fellowship.

References

Aizen, J., Huttenlocher, D., Kleinberg, J., & Novak, A. (2004). Traffic-based feedback on the web. *Proc. Natl. Acad. Sci.*, *101*, 5254–5260.

Allan, J., Papka, R., & Lavrenko, V. (1998). On-line new event detection and tracking. *SIGIR '98.*

Blei, D., & Lafferty, J. (2005). Correlated topic models. *NIPS '05.*

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *JMLR.*

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *J. of the Am. Soc. of Inf. Sci.*, *41.*

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, *29.*

Ghahramani, Z., & Jordan, M. I. (1995). Factorial hidden Markov models. *NIPS '95.*

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *KDD '03.*

Krause, A., & Guestrin, C. (2005). Optimal nonmyopic value of information in graphical models - efficient algorithms and theoretical limits. *IJCAI.*

Krause, A., Leskovec, J., & Guestrin, C. (2006). *Data association for topic intensity tracking* (Technical Report CMU-ML-06-100). Carnegie Mellon University.

Lerner, U. (2002). *Hybrid bayesian networks for reasoning about complex systems.* Ph.d. thesis, Stanford University.

Lerner, U., & Parr, R. (2001). Inference in hybrid networks: Theoretical limits and practical algorithms. *UAI.*

Ng, B., Pfeffer, A., & Dearden, R. (2005). Continuous time particle filtering. *IJCAI.*

Nodelman, U., Shelton, C., & Koller, D. (2003). Learning continuous time bayesian networks. *UAI.*

Segal, R. B., & Kephart, J. O. (1999). Mailcat: an intelligent assistant for organizing e-mail. *AGENTS '99.*

Swan, R., & Allan, J. (2000). Automatic generation of overview timelines. *SIGIR '00.*

Trivedi, K. (2002). *Probability and statistics with reliability, queuing, and computer science applications.* Prentice Hall.

Yang, Y., Ault, T., Pierce, T., & Lattimer, C. W. (2000). Improving text categorization methods for event tracking. *SIGIR '00.*