

Continuous-time Bayesian modeling of clinical data

Sathyakama Sandilya & R. Bharat Rao

Computer-Aided Diagnosis & Therapy Group
Siemens Medical Solutions USA, Inc., Malvern, PA 19355

January 9, 2004

Abstract

Inference from hospital patient records is difficult because data collection is done at arbitrary (not evenly-spaced) time intervals, and key clinical information is recorded only in unstructured form (as free text in doctors' notes). We present REMIND, a framework for performing inference from patient records based upon continuous-time Markov models and Bayesian networks. We empirically justify the need for such a complex model. REMIND only uses easily-available domain knowledge which we obtain from physicians and medical literature, which may be inaccurate. We also demonstrate the robustness of our approach to parameter assignment.

1 INTRODUCTION

Hospitals routinely collect patient data in multiple *data sources*. Structured data sources, such as financial, lab, and pharmacy databases, record patient information in tables. However, key clinical information is also stored in nonstructured data sources as free text (progress notes, radiology reports). Our research focuses on removing a key barrier: *the lack of structured high-quality clinical data, particularly reliably recorded outcomes, which makes patient records difficult to analyze.*

Consider a sample record for a cancer patient over a period of 2 years, who underwent chemotherapy for 4 months. This record may have 3 data sources, two structured (the drug and billing database) and one unstructured (transcribed doctors' dictations). Each data source records data at arbitrary points in time, not at evenly spaced time intervals (e.g., doctor's visits are not evenly spaced, chemo only occurs for a brief time). Sources also have differing degrees of reliability. The structured data repositories are typically missing critical clinical information (e.g., about outcomes), and may also contain errors. For instance, the ICD-9 diagnosis codes used to bill insurance companies reflect financial, rather than clinical, priorities. These are notoriously

inaccurate from a *clinical point of view* and are not very useful for clinical analysis (with error rates $\sim 30\%$ [5]). The free text in a transcribed doctors' dictation, may be reliable when read and understood by another doctor. However, to perform automated inference, information extracted from the text (via natural language processing) may be unreliable; typically due to uncertainties in the extraction, and sometimes due to errors in the original text.

Thus, any system that performs inference from existing patient records must deal with unreliable data gathered at arbitrary points in time. Further, as we would wish to rapidly deploy this system for multiple diseases in different hospitals, any clinical domain knowledge required must be modular, simple, and easy-to-acquire.

To perform inference from patient records, we model the evolution of the patient's disease state as a continuous-time Markov random process from which we obtain temporally nonuniform samples. We motivate the need for a dynamic model for the evolution of patients' disease state and empirically demonstrate the need to account for the nonuniform temporal sampling of the process.

Section 2 briefly reviews the REMIND (for Reliable Extraction and Meaningful Inference from Nonstructured Data) framework (for details see [9]). Section 3 describes the application domain and our study population: 344 colon cancer patients at NEORCC, a cancer care center in Ontario. NEORCC does not reliably record information about the most crucial of outcomes, recurrence of cancer. Section 4 describes the experimental setup. "Ground truth" is established by a colon cancer specialist after retrospective review of all 344 patient charts. Initial examination of the cases where REMIND's results differed from the specialist's has turned up 2 patients, where the specialist changed his retrospective diagnosis of recurrence. We empirically demonstrate the need for the model structure chosen in REMIND, by measuring the performance on this data set. Our second set of experiments analyze the robustness of REMIND's

conclusions to systematic variations in disease-specific knowledge used by REMIND. Section 5 concludes with a summary of this paper.

2 APPROACH

2.1 Problem Definition Let the state of the system, S , be a continuous-time random process taking values in a finite set Σ . Let $T = \{t_1, t_2, \dots, t_n\}$, where $t_i < t_{i+1}$, be the n “times of interest” when S has to be estimated. Let S_i refer to the sample of S at time $t_i \in T$. Let V be the set of variables of interest that depend upon S . Let O be set of all (probabilistic) observations for all variables, $v \in V$. Let O_i be the set of all observations “assigned” to t_i , i.e., all observations about variables in V , that are relevant for this time-step t_i . Similarly, let $O_i^k(v)$ be the k^{th} observation for variable v “assigned” to t_i . Let \mathbf{S} be a random variable in Σ^n ; i.e., each realization of \mathbf{S} is a sequence (S_1, S_2, \dots, S_n) .

GOAL: Estimate optimally a given set of variables for the patient $v_i(t)$ at given times of interest.

The specific instance of this problem that we present results for is the estimation of the sequence of disease states of the patient, i.e., estimation of \mathbf{S}_{MAP} , the maximum *a posteriori* estimate of \mathbf{S} given O .

2.2 Model We model S as a continuous time Markov process (from which we observe non-uniform samples). For simplicity, we also assume that S is time-invariant. We model the probabilistic relationships among $\{S_t, v_1(t), v_2(t), \dots, v_k(t)\}$ using a Bayesian Network. We assume that given S_t , for any $1 \leq i, j \leq |V|$ and $t' \neq t$, $v_i(t)$ is independent of $v_j(t')$ and $S_{t'}$.

A continuous-time Markov random process is characterized by the distribution of the amount of time the system spends in any state, and the probabilities with which it makes a transition to a state j given that it leaves state i , $i \neq j$. We assume that the *dwell time* distribution is exponential. Hence the parameter of interest there is the mean dwell time, i.e., $P[D_i > t] = \exp(-\lambda_i t)$, where D_i is the dwell time in state i whose mean is $1/\lambda_i$. Given that the system leaves state i , it goes to state j with probability q_{ij} .

2.3 Implementation REMIND first extracts information about the variables, $v \in V$ from the data sources, and converts them into a uniform representation. Each o_i is drawn entirely from a single piece of information in a data source (e.g., from a phrase in a sentence, or a row in a database table), and hence is assumed to be inherently undependable. The observations are represented as *a posteriori* distributions of variables, and converted into likelihood findings for computation.

The above conditional independence assumptions

allow us to split this into two steps. First of these is combination of observations at a fixed point in time, and the second is the propagation of these inferences across time. For each time instant of interest, we use the Bayesian Network connecting the variables at that time to compute the posterior distributions of all the variables at that time given all the observations at that time. For inference across time, we may now use a standard dynamic programming based approach (e.g. the Viterbi algorithm, see [8] for details). Further details on the implementation can be found in [9].

3 COLON CANCER APPLICATION

We consider a retrospective study of 344 Stage III colon cancer patients at NEORCC, a cancer care hospital in Sudbury, Canada. The principal source of structured patient data is OPIS, an oncology patient relational database. OPIS contains data about patient demographics, staging, diagnosis date, and administration of drugs. However, OPIS contains no reliable information about the most important outcome for these patients – recurrence of cancer. The principal source of unstructured patient data is another database that stores “doctors’ dictations” (transcribed free text). Patients averaged 10.8 dictations, with a maximum of 53 dictations. 18 patients had just 1 dictation, and 54 patients had 3 or fewer dictations. (As indicated earlier, T and n vary for different “realizations of the process”, i.e., patients.)

The domain knowledge requirements for REMIND are: identify state S (cancer has recurred or not) and the other variables of interest V , define extraction information, define T for each patient, and a mechanism to assign observations to $t \in T$, identify probabilistic dependencies, and define the possible state transitions. For this problem, the variables of interest as suggested by oncologists are the patient’s disease state (recurrent/non-recurrent), CEA – a tumor marker test (numerical, but converted to low/high), whether or not chemotherapy was administered (true/false), the if administered, the chemotherapy intent (adjuvant/palliative). The network for a particular instant in time is a tree with the disease state at the root, and the other variables being the children of the state. Each variable, however, may have multiple findings at the same time instant. Due to the conditional independence assumption we have made, the network connecting all the findings for a patient across time is also a tree as the state at one point in time is a child of the state at the previous time instant. These modeling assumptions make it feasible for us to perform exact inference for this particular problem. All the variables listed above are binary-valued. The disease state evolution is constrained, because a patient whose cancer has recurred

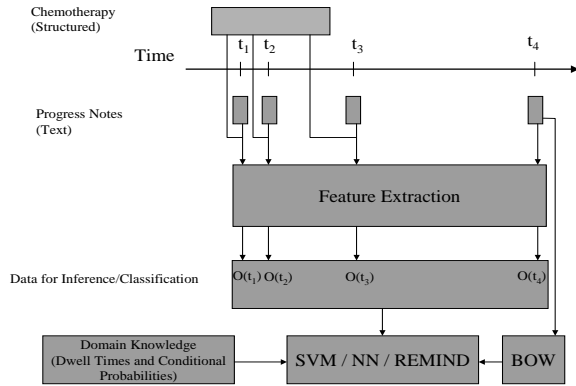


Figure 1: Data Flow

cannot subsequently become non-recurrent, which in turn, implies that the dwell-time in the recurrent state is ∞ . Hence, the model parameters that are of importance in the inference process are *extraction confidence*, *dwell time* in the “non-recurrent” state, and the distributions in the Bayesian network.

4 EXPERIMENTAL RESULTS

The key outcome (S) recurrence is only recorded at doctor’s visits. The performance task is to classify correctly the patient population as having recurred or not (equivalent to the value of S_n at the final state). Moreover, we would also like to know when each patient recurred, which is equivalent to classifying each patient visit (i.e., each S_i) as being one in which the patient is recurrent or not while accounting for temporal constraints on the process. Figure 1 illustrates the data flow model. Data for a single patient with doctor’s visits at $T = \{t_1, t_2, \dots, t_n\}$ is converted into a set of time-stamped observations O_i for each t_i . The feature extraction framework uses phrase spotting rules to produce observations with an *extraction confidence* which encodes the reliability with which this data is stored in the source, and the reliability of our extraction process. REMIND subsequently converts these observations into likelihood findings for the Bayesian Network. The distribution in each observation extracted from the data source is a function of the prior probability of the variable and the *extraction confidence*. As we want to demonstrate that REMIND’s parameters are robust and can be easily acquired without fine-tuning, we currently attach the same *extraction confidence* to all observations.

The phrase spotting rules (and other REMIND parameters) are assigned in consultation with domain experts, from the literature, and on review of a random

40% of the patient records. Our notion of a test set (the remaining 60%) is identical to that in standard machine learning; however, no cross-validation is possible with different training sets. To ensure uniformity, the other classifiers are also trained with the same 40% of the records and are tested on the remaining 60%. In this section, we empirically demonstrate the value of domain knowledge, continuous time temporal modeling, and the insensitivity of our framework to parameter assignment.

4.1 Domain Knowledge & Temporal Modeling

Table 1 compares REMIND’s performance against support vector machines (SVM), nearest neighbor (NN), and naive Bayes (NB). All algorithms use the probabilistic observations generated from the free text, converted into a 68-feature vector as input. In addition, we used bag-of-words (BOW) classifiers [7] whose input is the entire text document. Table 1 shows that REMIND clearly outperforms these systems by using easily available domain knowledge and temporal constraints. (This seems to load the dice in favor of REMIND; however, the difficulty of getting data with ground truth – our expert took 3 months to classify these 344 records – makes learning unreliable. Hence, we need to include domain knowledge in the process.) As a final step, instead of using the phrase-spotting rules, we used the single-document classifications produced by BOW to generate the O_t for REMIND. Although the results are not as good as those with the manually-assigned parameters, they are very promising and indicate that we may be able to learn many of REMIND’s parameters from small amounts of data. Table 1 also presents results of using these classification algorithms to estimate the final state of each patient.

However, it is still not clear that the added complexity of a continuous-time model is necessary. Table 2 presents the performance of REMIND with 3 different temporal evolution models. As no training is involved in these experiments, we present classification results on the entire dataset of 3687 patient visits.

IID: In this case, there is no temporal reasoning – the state at each time of interest is modeled as being independent of previous states, and has the same distribution. Note that there is no means of imposing temporal constraints on the disease (such as, a recurrent patient cannot subsequently become non-recurrent).

Discrete-Time Markov: This model assumes a temporal evolution, but ignores inter-visit time, i.e., the transition probabilities are independent of the lapsed time between visits. This model can encode temporal constraints, but effectively models the process as being sampled uniformly.

Continuous-Time Markov: This is the model used

| Algorithm | Visit Classification | | | | | Patient Classification | | | | |
|-----------|----------------------|-----|-----|------|------|------------------------|----|------|------|------|
| | FP | FN | FP% | FN% | Er% | FP | FN | FP% | FN% | Er% |
| NB | 50 | 231 | 2.9 | 48.1 | 12.9 | 19 | 3 | 12.8 | 5.2 | 10.7 |
| SVM | 32 | 273 | 1.9 | 56.9 | 14.0 | 6 | 17 | 4.0 | 29.3 | 11.2 |
| 3-NN | 67 | 284 | 3.9 | 59.2 | 16.1 | 2 | 26 | 1.3 | 44.8 | 15.6 |
| BOW | 126 | 166 | 7.4 | 34.6 | 13.4 | 23 | 18 | 15.5 | 31.0 | 19.9 |
| BOW+REM | 67 | 122 | 3.9 | 25.4 | 8.7 | 5 | 17 | 3.4 | 29.3 | 10.7 |
| REMIND | 102 | 40 | 6.0 | 8.3 | 6.5 | 5 | 10 | 3.4 | 17.2 | 7.3 |

Table 1: Classifier Comparison on test data set (206 patients, 2181 visits)

| Model for State Evolution | Dec. Rec. | FP | Dec. Nonrec. | FN | FP% | FN% | Error% / visit |
|---------------------------|------------|------------|--------------|-----------|-------------|-------------|----------------|
| I.I.D | 859 | 75 | 2828 | 357 | 2.6% | 42.4% | 11.7% |
| Discrete Time | 1468 | 758 | 2219 | 121 | 26.6% | 14.4% | 23.8% |
| Continuous time | 948 | 168 | 2739 | 61 | 5.9% | 7.3% | 6.2% |

Table 2: Different state evolution models for estimating disease state for 3687 visits

by REMIND as discussed above and in Section 2.3.

The first two models are obviously coarser than that used in REMIND. Continuous-time temporal modeling reduces errors by almost a factor of 2, which demonstrates that to solve inference problems with clinical data, the complexity of continuous-time modeling is justified, and indeed, necessary.

4.2 Robustness to parameter variation As mentioned before, the model parameters used in REMIND have been obtained from medical literature, or in conversations with domain experts. As these are likely to be incorrect, we wished to study the effect of variation of these parameters on the performance of REMIND. Figure 2 (a) shows that the degradation in performance is very graceful as the *extraction confidence* is varied from 0.1 to 0.9. Figure 2 (b) presents the performance variation when we vary the *dwelt time* parameter from 5 months to 1000 months. Note that the performance of REMIND is stable all the way from 20 to 1000 months. (The REMIND experiments in Tables 1&2 use *extraction confidence* = 0.5 and *dwelt time* = 50 months.)

We also vary the conditional probability distributions of the variables in our models by adding log-normal noise to the odds as in [3]. Zero mean random noise with varying variance is added to the odds in each conditional distribution. The effect of adding log-odds noise to the distribution of a boolean variable with probability 0.8 is depicted in Figure 3 (a) for $\{\sigma = 0, 0.2, 0.5, 1, 1.5, 2\}$. As we can see from this plot, adding noise with a large standard deviation often

changes the distribution drastically. Figure 3 (b) shows that the inference engine is remarkably robust to such changes for different performance measures (visit and patient level sensitivity/specificity and exact sequence prediction).

5 DISCUSSION

Hospital patient records are a valuable source for outcomes analysis and data mining. Unfortunately, most of the important clinical data is present in unstructured format in free text doctor’s dictations, and as such is not readily amenable to analysis. Analyzing these already-collected patient records can, however, lead to important clinical insights that can be subsequently verified in a prospective trial, or to improve standards of care at medical institutions. We have proposed REMIND, a general framework for performing Bayesian inference on random processes that are sampled at arbitrary time instants by integrating information from structured and unstructured parts of a patient record. We have applied our technique to extract recurrence, a complex time-varying outcome, for colon cancer patients (and for cardiac patients [9]). By automatically structuring already collected patient records, we make existing patient data available for analysis.

We have justified the structure and assumptions of our model by comparing its performance against that of traditional models for inference. Our technique performs well for two important reasons: first, the temporal reasoning ensures that we take into consideration the hints about the patient’s present state that are present

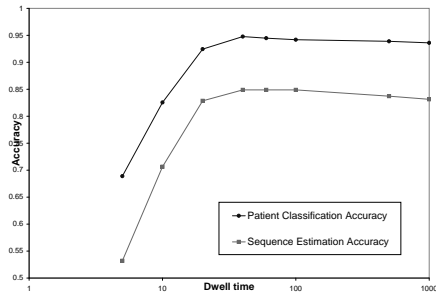
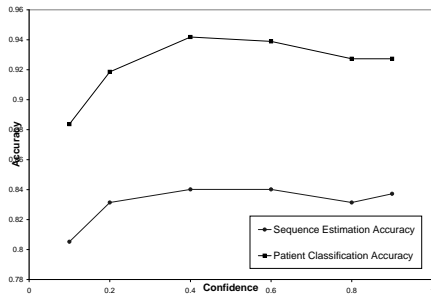


Figure 2: Accuracy Variation with (a) Confidence and (b) Dwell time

in data from the past and the future, and second, the continuous-time state evolution model ensures that we account for the non-uniformity in temporal sampling. We have also analyzed the performance of REMIND under different assignments and demonstrated that it is very robust to parameter variations.

As we use Bayesian models and estimation, our work draws heavily on earlier work on Bayesian networks and graphical models (see [4] for an overview). Also related to the results we present in this paper is the discussion of the insensitivity of diagnosis based on Bayesian models to parameter settings [3]. REMIND can clearly benefit from using better Natural Language Processing methods [6] as a significant fraction of the observations are gathered from text. Augmenting the domain knowledge provided to REMIND with a general lexical reference [1] or a medical language dictionary (SNOMED) should improve performance.

References

- [1] Fellbaum, C., *WordNet: An Electronic Lexical Database*. MIT Press, May 1998.
- [2] Heckerman, D. A tutorial on learning with Bayesian

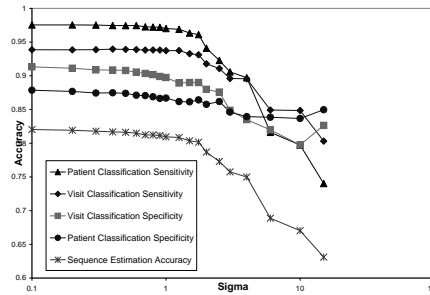
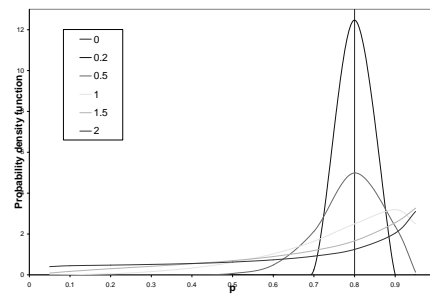


Figure 3: (a) Effect of log-odds noise on distribution and (b) Accuracy Variation with log-odds noise

networks. *Microsoft Research tech. report*, MSR-TR-95-06, 1996.

- [3] Henrion, M., Pradhan, M., Del Favero, B., Huang, K., Provan, G., O'Rorke, P., *Why is diagnosis using belief networks insensitive to imprecision in probabilities?*, Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence, pp. 307-14.
- [4] Jensen, F.V. *An introduction to Bayesian Networks*. UCL Press, 1996.
- [5] Jollis J.G., Ancukiewicz M., DeLong E.R., et al., *Discordance of databases designed for claims payment versus clinical information systems*, Annals of Internal Medicine 1993;119(8):844-850.
- [6] Manning, C.D., Schütze, H. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts, MIT Press.
- [7] McCallum, A., Freitag, D., Pereira, F. Maximum Entropy Markov Models for Information Extraction and Segmentation. *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*.
- [8] Rabiner, R.L. A Tutorial on Hidden Markov Models and Selected Applications in Speech recognition. *Proceedings of the IEEE*, 77(2), 257-286.
- [9] Rao, R.B., Sandilya, S., Niculescu, R., Germond, C., Rao, R.H., Clinical and Financial Outcomes Analysis with Existing Hospital Patient Records, KDD 2003.