Probabilistic Abstraction of Multiple Longitudinal Electronic Medical Records

Michael Ramati and Yuval Shahar

Medical Informatics Research Center, Department of Information Engineering,
Ben-Gurion University, P.O.B. 653, 84105 Beer-Sheva, Israel
{ramatim, yshahar}@bgu.ac.il

Abstract. Several systems have been designed to reason about longitudinal patient data in terms of abstract, clinically meaningful concepts derived from raw time-stamped clinical data. However, current approaches are limited by their treatment of missing data and of the inherent uncertainty that typically underlie clinical raw data. Furthermore, most approaches have generally focused on a single patient. We have designed a new probability-oriented methodology to overcome these conceptual and computational limitations. The new method includes also a practical parallel computational model that is geared specifically for implementing our probabilistic approach in the case of abstraction of a large number of electronic medical records.

1 Introduction

The commonly occurring task of Temporal Abstraction (TA) was originally defined as the problem of converting a series of time-oriented raw data (e.g., a time-stamped series of chemotherapy-administration events and various hematological laboratory tests) into interval-based higher-level concepts (e.g., a pattern of bone-marrow toxicity grades specific to a particular chemotherapy-related context) [1]. Former solutions [1-4], although being evaluated as fruitful, maintained several unsolved subproblems. These subproblems seem common to some of other methods suggested for solving the TA task as well as closely related systems applied in the clinical domain (e.g., [5-7]). Thus, Considering these challenging subproblems suggests an additional method.

At least three subproblems in the former methods can be pointed out, which we propose to solve through the method discussed in this paper. First, raw clinical data, to which the temporal reasoning is being applied, are assumed as certain – that is, typically no mechanism is suggested for handling the inherent impreciseness of the laboratory tests taken to obtain the clinical data. Second, current mechanisms used for completing missing data in an electronic medical record are typically not sound and are incomplete. For example, in the case of the KBTA method, a knowledge-based interpolation mechanism is used [8]. However, completion of missing values is supported only for bridging gaps between two intervals, in which the proposition (e.g., anemia level) had the same value (e.g., moderate anemia). Furthermore, the value concluded by inference is too crisp, and a threshold is used for computing it with absolute certainty, eliminating uncertainty and leading to potentially unsound conclu-

S. Miksch et al. (Eds.): AIME 2005, LNAI 3581, pp. 43 – 47, 2005.

sions. Third, no special mechanism has been devised for multiple patient abstraction. That is, so far temporal abstraction was performed on a single patient only.

The proposed method, *Probabilistic Temporal Abstraction* (PTA), decomposes the temporal abstraction task into three subtasks, that solve the case of a single patient, and two more subtasks that solve the case of multiple patients.

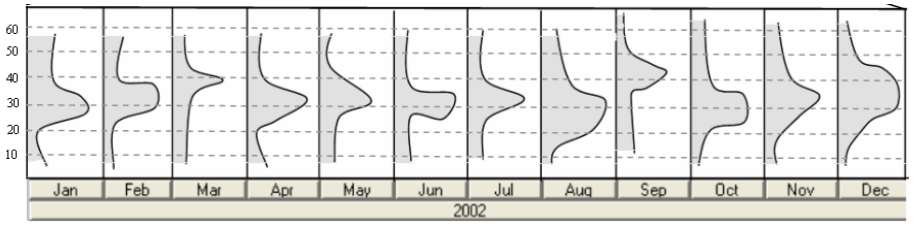

Fig. 1. A typical instance of using the PTA method: the value (*vertical axis*) distribution of a certain medical concept appears for different (in this case consecutive) periods along the time axis. The medical concept, which can be either raw or abstract, and the specification of the set of periods (including the time granularity) are determined by the application using the PTA method

2 The PTA Method

The main computational concept in our methodology is the PTA chain. A *PTA chain* is defined as the application of any subset of the following composition of subtasks, while preserving the relative order among them:

$$Coarsen \circ Correlate \circ Aggregate \circ Transform \circ Interpolate(data). \quad (1)$$

These subtasks are introduced below using the following known notions in probability theory. A *stochastic process* {X(t): t in T} is a set of random variables, and may represent a clinical observation, a medical intervention, or an interpretation context of some clinical protocol. The index is often interpreted as time, and thus X(t) is referred as the *state* of the process at time t. The set T is called the *index set* of the process.

2.1 The PTA Property and Temporal Interpolation

The central property of the PTA method is based on the notion of *temporal field*, as explicated below. Following this definition, the property states, that each unobserved state of some stochastic process is a linear combination of the temporal fields of the observed states of the process. Thus, the unobserved distribution of bone-marrow toxicity grades is a linear combination of all of the observed distributions, before and after it. An observed state is distributed as a function of the reliability or precision of the clinical test taken (variance) and the value sampled (mean), and induces a *field*[1] over its temporal environment, expressing the temporal knowledge about the stochastic process in question, such as a *periodic* behavior, *monotonic* change, or *persistence*.

[1] In the sense of an electromagnetic field.

For example, suppose a stochastic process with a periodic behavior and cycle length *c*. The temporal field of an observed state of such stochastic process could be as follows:

$$(field_{\vec{X}}(t_s))(t_i) = \sin\left(\frac{\pi}{c}\cdot(|t_i - t_s| \bmod c)\right)\cdot X_{t_s}. \quad (2)$$

Having multiple fields induced over an unobserved state necessitates the use of weights, to express the notion that the closer-in-time the observed state is – the more relevant it is. Therefore, there is a need to choose a monotonic decreasing function of absolute time differences between a dependent state and its inducing observed states. A natural choice for the monotonic decreasing weighting function would be a normal density, where its variance (σ^2) determines the temporal tolerance of observed states of the stochastic process. Thus, *w* may hold:

$$w_{\vec{X}}(\Delta t) = f_W(\Delta t) \quad , W \sim Normal(0, \sigma^2). \quad (3)$$

The *Temporal Interpolation* subtask is aimed at estimating the distribution of a stochastic process state, given the distributions of some of its other states. For example, estimating the distribution of raw hematological data or derived concepts (such as bone-marrow toxicity grades) during a week in which raw data were not measured, using the distribution of values before and after that week. Applying the interpolation subtask does not increase the abstraction level of the underlying stochastic process, but rather serves the role of a core operation that enables the application of actual temporal abstraction. The subtask of interpolation is solved by the application of the PTA property. Thus, the subset of sampled states which participate in the calculation of each unobserved state determines the precision of its distribution, and could be determined given the temporal weighting function. If we interpolate in t_i and know all sampled values t_s (where s=-∞ stands for prior distribution), then:

$$X_{t_i} = \frac{1}{\sum_{t_s} w_{\vec{X}}(t_i - t_s)} \sum_{t_s} w_{\vec{X}}(t_i - t_s)\cdot(field_{\vec{X}}(t_s))(t_i). \quad (4)$$

For the case in which updates to the underlying clinical data occur, we consider a hierarchical system of states, where each unobserved state has a set of observed parent states, as depicted by Pearl [9]. In case the sample is updated, propagating the new piece of evidence we are viewing as the perturbation that propagated through a Bayesian network via message-passing between neighboring processors.

2.2 Other Subtasks and Their Mechanisms

Temporal abstraction for a single patient requires one basic subtask, temporal interpolation, and the two interpolation-dependent subtasks explicated below. The *Temporal Coarsening* subtask is aimed at the calculation of a stochastic process at a coarser time granularity, according to the following formula:

$$X_{[t_i,t_j]}=\frac{1}{j-i+1}\cdot\sum_{k=i}^{j}X_{t_k}. \tag{5}$$

The *Temporal Transformation* subtask is aimed at the generation of a stochastic process, given stochastic processes of a lower abstraction level, according to the following formula:

$$Y_t=(g(\vec{X}_1,\ldots,\vec{X}_n))(t). \tag{6}$$

For example, deriving bone-marrow toxicity grade distribution, given the distributions of the raw white blood cell and platelet counts. A special transformation function is:

$$(change(\vec{X}))(t_i)=X_{t_i}-X_{t_i-1}. \tag{7}$$

Applying the TA task to multiple patients requires extra subtasks, such as the ones explicated below. However, these subtasks fit also sophisticated needs of abstraction for a single patient. The *Temporal Aggregation* subtask is aimed at the application of an aggregation function (such as minimum, maximum, average, etc.) to temporally corresponding states of stochastic processes of the same sample space and independent patients. In the case of a single patient, the aggregated states are taken from the same process. The *Temporal Correlation* subtask is intended to mainly compare two patient populations, but should work the same when comparing different time periods of the same patient, resulting in a series of correlation factors between corresponding states of the given stochastic processes. An example for a single patient would be the contemporaneous correlations between height and weight or correlation of height during different periods for the same person.

3 Parallel Implementation

The computational model used to compute a PTA chain is *goal-driven, bottom-up* and *knowledge-based*. That is, the main algorithm is required to compute the result of a PTA chain (the goal), given the transformation and interpolation functions (the temporal knowledge) as well as the access to the clinical data, beginning at the raw (lowest abstraction level) clinical data. The computational model is parallelized (and hence *scalable* [10]) in three orthogonal aspects: (1) Time, during the calculation of the PTA chains' states; (2) Transformation, during the calculation of the transformation arguments; and (3) Patient, during the calculation of the PTA chains for multiple patients.

The PTA architecture is in the process of being fully implemented using is the C++ programming language, the Standard Template Library (STL), and the MPICH2 implementation of the Message-Passing Interface (MPI)[2], an international parallel programming standard. The implementation is thus object-oriented and platform-independent. The implementation is in the process being integrated into the IDAN system [11], which satisfies the need to access medical knowledge and clinical data sources.

[2] http://www.mpi-forum.org/

4 Discussion

The new probabilistic method has removed several limitations of former methods. First, the use of PTA chains enables the expression of uncertainty in the underlying clinical data. Second, two mechanisms were developed for temporal abstraction of the clinical data of multiple patients. Third, the interpolation mechanism was shown to be sound and complete. However, observed clinical data are assumed to be independently distributed. This assumption could be easily removed, given the necessary domain-specific conditional distribution functions.

The Markovian property (i.e., the conditional distribution of any future state, given the present state and all past states, depends only on the present state) is not assumed by the PTA method, where past states may be relevant in computing future states. The interpolation in the PTA model is performed at the lowest abstraction level only, as opposed to being repeatedly performed at every abstraction level as in the KBTA method [1]. Finally, the components of the PTA method are highly modular and do not assume, for example, a particular temporal representation.

References

1. Y. Shahar: A Framework for Knowledge-Based Temporal Abstraction. Artificial Intelligence (1997) 90:79-133
2. M.J. O'Connor, W.E. Grosso, S.W. Tu, M.A. Musen: RASTA: A Distributed Temporal Abstraction System to facilitate Knowledge-Driven Monitoring of Clinical Databases. MedInfo, London (2001)
3. A. Spokoiny, Y. Shahar: A Knowledge-based Time-oriented Active Database Approach for Intelligent Abstraction, Querying and Continuous Monitoring of Clinical Data. MedInfo (2004) 84-88
4. M. Balaban, D. Boaz, Y. Shahar: Applying Temporal Abstraction in Medical Information Systems. Annals of Mathematics, Computing & Teleinformatics (2004) 1(1):54-62
5. M.G. Kahn: Combining physiologic models and symbolic methods to interpret time varying patient data. Methods of Information in Medicine (1991) 30(3):167-178
6. I.J. Haimowitz, I.S. Kohane: Automated trend detection with alternate temporal hypotheses. Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann, San Mateo (1993) 146-151
7. A. Salatian, J. Hunter: Deriving trends in historical and real-time continuously sampled medical data. Journal of intelligent information systems (1999) 13:47-71
8. Y. Shahar: Knowledge-Based Temporal Interpolation. Journal of Experimental and Theoretical Artificial Intelligence (1999) 11:123-144
9. J. Pearl: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers (1987)
10. K. Hwang, Z. Xu: Scalable Parallel Computing, WCB McGraw-Hill (1998)
11. D. Boaz, Y. Shahar: A Framework for Distributed Mediation of Temporal-Abstraction Queries to Clinical Databases: Artificial Intelligence in Medicine (in press)