

Integration of Temporal Abstraction and Dynamic Bayesian Networks in Clinical Systems. A preliminary approach

Kalia Orphanou¹, Elpida Keravnou², and Joseph Moutiris³

- 1 Department of Computer Science, University of Cyprus, Nicosia, Cyprus
korfan01@cs.ucy.ac.cy
- 2 Cyprus University of Technology, Limassol, Cyprus rector@cut.ac.cy
- 3 Department of Cardiology, Nicosia, General Hospital, Cyprus
moutiris@ucy.ac.cy

Abstract

Abstraction of temporal data (TA) aims to abstract time-points into higher-level interval concepts and to detect significant trends in both low-level data and abstract concepts. TA methods are used for summarizing and interpreting clinical data. Dynamic Bayesian Networks (DBNs) are temporal probabilistic graphical models which can be used to represent knowledge about uncertain temporal relationships between events and state changes during time. In clinical systems, they were introduced to encode and use the domain knowledge acquired from human experts to perform decision support.

A hypothesis that this study plans to investigate is whether temporal abstraction methods can be effectively integrated with DBNs in the context of medical decision-support systems. A preliminary approach is presented where a DBN model is constructed for prognosis of the risk for coronary artery disease (CAD) based on its risk factors and using as test bed a dataset that was collected after monitoring patients who had positive history of cardiovascular disease. The technical objectives of this study are to examine how DBNs will represent the abstracted data in order to construct the prognostic model and whether the retrieved rules from the model can be used for generating more complex abstractions.

1998 ACM Subject Classification I.2.1 Applications and Expert Systems

Keywords and phrases temporal abstraction, medical prognostic models, dynamic Bayesian network, coronary artery disease

Digital Object Identifier 10.4230/OASlcs.ICCSW.2012.102

1 Introduction

Advances in the field of Artificial Intelligence led to the development of intelligent clinical data analysis systems, that are designed to provide computer-based support in medical tasks by automating, for example, diagnostic reasoning. In general, the purpose of medical data analysis systems is to aid care providers reach the best possible decisions for any patient, to help them understand what the possible consequences of their decisions/actions are and if necessary to take corrective actions in a short time interval.

Temporal abstraction (TA) [16] abstracts time-point based data into higher-level, interval based concepts under a given context. Abstraction of time-oriented clinical data aims to close the gap between general medical knowledge and specific patient data. Medical knowledge is expressed in a general form (association rules, patient management protocols) while patient data are specific (history of patients, results of laboratory and physical examinations). The



© Kalia Orphanou, Elpida Keravnou, and Joseph Moutiris;
licensed under Creative Commons License NC-ND
2012 Imperial College Computing Student Workshop (ICCSW'12).
Editor: Andrew V. Jones; pp. 102–108



OpenAccess Series in Informatics

OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

ICCSW

derived abstracted concepts are useful for different tasks such as decision making, therapy planning and summarization of a patient's record.

According to Bellazzi [1], time-stamped entities are called events and their abstract representation, given by TAs as sequences of intervals, are called episodes. The TA task can be divided into two subtasks, basic and complex abstractions. Basic temporal abstractions are state and trend abstractions which abstract events (time-stamped data) within episodes. Complex temporal abstractions abstract episodes (intervals) into other episodes. The aim of state abstraction is to derive maximal intervals over which there is no change in the state of some parameter. The aim of trend abstraction is to derive the significant changes and the rates of change in the progression of some parameters e.g stable, increasing, decreasing.

A Bayesian Network (BN) is an acyclic graph that represents a joint probability distribution over a set of random variables. It consists of two components, a directed acyclic graph and a probability distribution. Nodes on the graph represent the variables and edges represent the direct dependencies between the variables. BNs can make predictions or give explanations by computing the conditional probability table of each variable. BNs have been introduced as a knowledge representation method to encode and use the domain knowledge acquired from human experts in automated reasoning systems to perform diagnostic, predictive and explanatory tasks. They can represent knowledge even in cases of missing data or uncertain information.

Dynamic Bayesian Networks (DBNs) [3] are a temporal extension of standard BNs that are able to model stochastic processes. They utilize a representation of a dynamic process via a set of stochastic variables in a sequence of time-slices. More precisely, a DBN is a network with the repeated structure of a BN for each time slice over a certain interval. Consequently, a DBN is a tuple $(B1, B2)$, where $B1$ is a Bayesian Network that represents the prior distribution for the variables in the first time slice and $B2$ represents the transition model for the variables in two consecutive time slices. Relations between variables are divided into two types: transitional relations represent dependencies between variables between different time slices and local relations represent dependencies among variables in the same time slice. DBNs are usually assumed that they use the Markovian property: conditional probability distribution of each variable at time t , for all $t > 1$, depends only on the parents from the same time slice or from the previous time slice but not from earlier time slices.

This paper is organized as follows. Section 2 describes the preliminary approach of integrating these two areas under the domain of CAD. Section 3 describes related work. Section 4 discusses the potential advantages of this integration and future improvements of this approach and the paper ends with closing remarks.

2 Preliminary Approach

2.1 Data Description

Coronary artery disease (CAD) is one of the major causes of disability in adults as well as one of the main causes of death in the developed countries. The dataset was collected after 4 years of monitoring patients (2009–2012) under the supervision of the participating cardiologist (Dr. J. Moutiris) at the Cardiology Clinics of Nicosia General Hospital, Cyprus. The target group consists of 176 patients of 32–89 years old. The dataset includes physical and biochemical examination results. Some patients took physical and biochemical examinations every 5–6 months, some others took examinations once a year and some took three examinations over the four years period.

The principal goal of the dataset collection was to identify the risk factors of coronary

artery disease and their impact on patients' health. Risk factors are defined as: smoking, diabetes and arterial hypertension considering systolic and diastolic blood pressure, high levels of cholesterol, high levels of triglycerides, low HDL, high LDL and overweight. The inclusion criteria to the study included the presence of at least one event, such as acute coronary syndrome (ACS), acute myocardial infarction (AMI), percutaneous coronary intervention (PCI) and coronary artery bypass graft surgery (CABG).

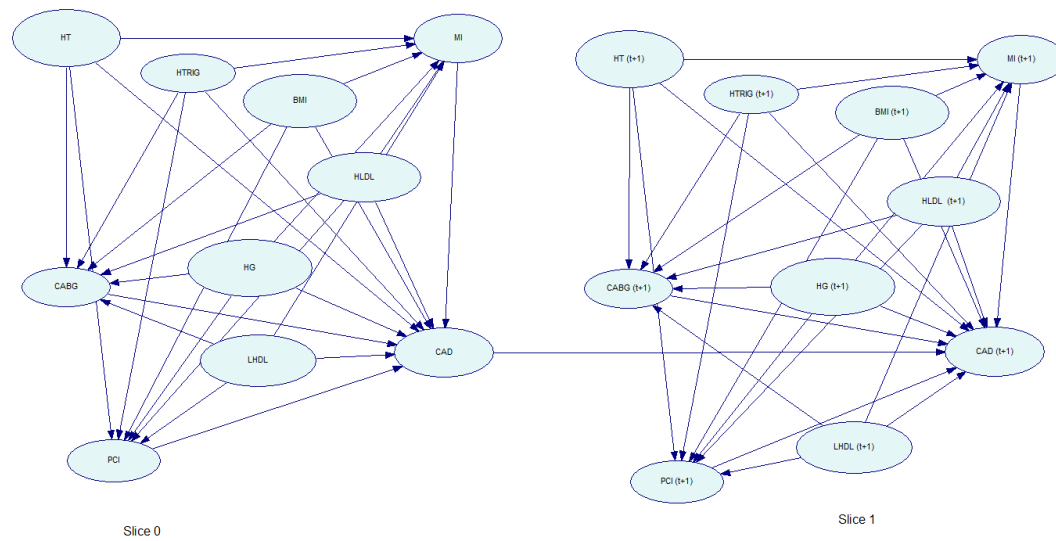
2.2 Methodology

The basic steps of the proposed methodology are:

1. Data cleaning which includes identifying fields, selecting variables for abstraction and coding data as shown in Table 1.
2. State and trend abstraction techniques will be used to abstract concepts based on an interval duration given by a domain (12-months). State abstractions will be generated based on domain expert knowledge whereas trend abstractions will be generated by extending the algorithm described in [14].
3. The derived state abstractions will be the new concepts which represent the presence or absence of risk factors during the 12-month period for each patient.
4. A DBN model will be constructed using three time slices and each time slice will represent the time period of 12 months (e.g. 01/01/2009–01/01/2010). Nodes represent the abstract concepts which are binary variables and edges represent the dependencies between the concepts through the same time-slice or through two consecutive time slices. The structure of the DBN (prior and transition model) representing the state abstraction concepts is based on domain expert knowledge and medical literature as shown in Fig. 1.
5. Expectation Maximization learning algorithm [10] will be used for learning parameters of the model using the abstracted data set. Bayesian Net Toolbox for Matlab [11] will be used for the construction of the DBN. The constructed model will be able to make predictions for the risk of coronary artery disease for a specific patient.
6. The junction tree inference algorithm [11] will be used to compute the prediction risk (probability). If the probability will be over 0.5 then the risk of the presence of cardiovascular disease is severe otherwise it is normal.

■ **Table 1** Variables for abstraction and coding data.

Variables	Code
Hypertension	HT
High Total Cholesterol	HTC
Low HDL	LHDL
High LDL	HLDL
High Triglycerides	HTR
High Glucose + Diabetes	DM
Smoking	SK
Overweight	BMI
Myocardial Infarction	MI
Acute Coronary Syndrome	ACS
Coronary Artery ByPass Graft	CABG
Percutaneous coronary intervention	PCI
Diet, Exercise	DIET, EX
Risk of CAD	hidden variable – CAD



■ **Figure 1** Coronary Artery Disease Model Structure Over Two Time Slices ($t=0$ and $t=1$): Nodes represent some of the variables as displayed in Table 1 and edges represent their dependencies (local and transitional relations). A transitional relation exists only with respect to the hidden node CAD.

2.3 Temporal Abstractions

State abstractions correspond to expressions like high blood pressure associated to a time interval in which such behavior occurs. The labels given to the associated intervals are: ‘normal levels’ (value = 1) or ‘abnormal levels’ (value=2) which are given by the domain expert as shown in Table 2. If a patient took more than one examination over the same year and the value of the state abstraction remains the same, the time points are joined into a maximal interval $[Iss, Ies]$. Alternatively, if a patient took more than one examination in the same year, but the state value of a variable is not the same for all examinations, then the following rules given by domain expert are applied:

- If a patient took two examinations during the desired time period and $V_s = 2$ (abnormal) at t_1 but $V_s = 1$ (normal) at t_2 then a risk factor is absent at $[t_1-t_2]$ e.g patient with high cholesterol during his/her first examination and cholesterol value decreases to normal levels from the first examination to the next one, then the risk of hypercholesterolaemia is absent during that period. Similarly, if the cholesterol value increases from normal (at t_1) to abnormal levels (at t_2) then the risk of hypercholesterolaemia is present.
- If a patient took three examinations during the desired time period and the value of state abstraction is the same during t_1 and t_2 then the presence of a risk factor depends on the value of the abstraction at t_1 and t_2 . For example, if $V_s = 2$ (abnormal) at t_1 and t_2 but $V_s = 1$ (normal) at t_3 then a risk factor is present during the period $[t_1 - t_3]$.
- If a patient took three examinations during the desired time period and the value of state abstraction is the same during t_2 and t_3 then the presence of a risk factor depends on the value of abstraction at t_2 and t_3 . For example, if $V_s = 1$ (normal) at t_1 but $V_s = 2$ (abnormal) at t_2 and t_3 the risk factor is present during the period $[t_1 - t_3]$.
- If a patient took three examinations during the desired time period but the values of state abstractions are different at all three time steps then the presence of risk factors depends only on the value of abstraction at t_3 . For example if $V_s = 2$ (abnormal) at t_1 , $V_s = 1$ (normal) at t_2 and $V_s = 2$ (abnormal) at t_3 then a risk factor is present during the period $[t_1 - t_3]$.

- As concerned the risk factor of hypertension which depends both on systolic and diastolic blood pressure, the above rules have to concern either systolic or diastolic blood pressure values.

Trend abstractions will be generated after applying a median filter to the dataset for removing noise. Then, the values of variables occurring at consecutive time points will be compared. Let us assume that t_1 is the time of the first consultation and t_2 is the time of the second consultation during the year. If the value of a variable at t_1 (V_1) is equal to the value of the variable at t_2 (V_2) then the value of the variable V (V_t) during the interval $[t_1 t_2]$ is 'steady'. If the value of a variable at t_1 (V_1) is less than the value of the variable at t_2 (V_2) then the value of the variable V (V_t) during the interval $[t_1 t_2]$ is 'increasing', otherwise, the value of the variable V (V_t) during the interval $[t_1 t_2]$ is 'decreasing'. A maximal interval where this behavior (state, increasing or decreasing) persists is derived.

- **Table 2** Categories of state abstractions as given by domain expert where each variable can take two possible state values depending on its raw value.

State Value	Normal for all patients	Normal for patients with Diabetes	Abnormal before the day of the event
Cholesterol (mg)	< 190	<170	
HDL (mg)	> 40	>40	
LDL (mg)	<100	<100	
Glucose (mg)	<110	<110	
BMI	<25	<25	
SBP (mmHg)	<130	<120	Hypertension = YES
DBP (mmHg)	<90	< 85	Hypertension = YES
Triglycerides (mg)	<150	<150	
Smoking	NO		EX (ex smoker)
Diet	YES		
Exercise	YES		

3 Related Work

Techniques/methods from these two areas are largely used independently of each other in many clinical domains and thus no specific integrations have been reported yet. Consequently, the goal of the proposed methodology is to combine temporal abstraction with Dynamic Bayesian networks and to apply this integration under the CAD clinical domain by developing a prediction model.

Several systems had been designed to abstract meaningful clinical concepts from raw clinical data such as TOPAZ, IDEFIX and VM, each one using its own abstraction methodologies [8, 5, 6]. The best known framework for temporal abstraction in clinical data, called Knowledge Based Temporal Abstraction (KBTA) was proposed by Shahar in [15] and implemented in many clinical domains. KBTA decomposes the temporal abstraction task into five computational subtasks, solved by corresponding temporal mechanisms. However, none of these approaches are able to deal with missing data and of the uncertainty that typically underlie clinical raw data. The KBTA method complete missing values only for bridging gaps between two intervals, in which the proposition (e.g., anemia level) had the same value (e.g., moderate anemia). The IDEFIX system can deal with uncertainty and also with missing data using the time-of-validity slot which specifies for how long the value of the attribute is considered valid. Moreover, when domain medical knowledge is not enough and prior probabilities about some disease are missing, the derived conclusions of IDEFIX

may not be valid. Ramati and Shahar [12] proposed a new methodology called Probabilistic Temporal Abstraction (PTA) to perform a temporal abstraction task to clinical raw data using a probabilistic approach. This approach is able to eliminate uncertainty and to deal with missing values.

Considerable work on dynamic models in medicine has been carried out by Leong and collaborators who have successfully used a combination of graphic models with Markov chains to solve problems in different medical domains such as head injury management [7], colorectal cancer management, neurosurgical intensive care unit monitoring and palate management [18]. Other applications of Dynamic Bayesian Networks in medicine include forecasting sleep apnea [4], management of patients with carcinoid tumor [17] and diagnosis and decision making after monitoring patients suffering from renal failure and treated by hemodialysis [13].

A classification/diagnosis model under the clinical domain of CAD had also been developed using decision trees in a previous work by Karaolis [9] using a similar dataset. Preprocessing of current dataset is based on this work. Related is also the work in reasoning over multiple levels of temporal granularity through Bayesian networks [2]. In this work, two approaches were proposed to incorporate temporal abstractions and explicitly represent complex temporal relationships using fluents and hierarchical Bayesian networks.

4 Conclusion and Future Work

The perceived advantages of combining TA with DBN are that this integration can handle incomplete evidence and uncertainty estimating disease outcomes which are usual problems in clinical systems and also represent current limitations in datasets. DBN can provide a concrete understanding of how causal dependencies and temporal precedence between abstract concepts influence a particular disease outcome. Moreover, this integration facilitates the main advantage of BNs which is their capability to integrate expert's knowledge with empirical data to model a disease.

Another advantage of integrating a DBN with TA is the ability to represent high-level abstracted concepts rather than low-level data, thus making the models simpler and at the same time more conceptual. Moreover, the interpretation of the prediction results will be context-based, allowing for more accurate decision making. In addition, uncertainty and errors in clinical datasets are very common and TA and DBN techniques can deal with such error-prone measurements.

In this paper, a preliminary approach of integrating temporal abstraction with DBN in CAD clinical domain is presented. The proposed approach has some limitations which will be overcoming at a later stage of this study. Although the dataset includes information about history of hypertension, diabetes, MI and ACS which may have occurred before 2009, the constructed model can only represent risk factors and events that hold during the period 2009–2012. The model should be extended in order to represent history of events as well. Another possible improvement is the developed model to be able to facilitate continuous improvement and innovation in medical tasks through incremental learning. This means that as new cases will be dynamically added to the network, the model must self-adapt its structure and adjust its parameters on-line. Furthermore, later stages of this study will work on how trend abstractions can be represented in a DBN and the possibility to develop an hierarchical model with abstracted data of different granularities.

References

- 1 Larizza C. Bellazzi R. and Riva A. Temporal abstractions for interpreting diabetic patients monitoring data. *Intelligent Data Analysis*, 2(1-4):97–122, 1998.
- 2 Brendan Burns and Clayton T. Morrison. Temporal abstraction in bayesian networks. In *AAAI Spring Symposium*, 2003.
- 3 T. Charitos. *Reasoning with dynamic networks in practice*. PhD thesis, Utrecht University, Netherlands, 2007.
- 4 Paul Dagum and Adam Galper. Forecasting sleep apnea with dynamic network models. In *Proceedings of the Proceedings of the Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-93)*, pages 64–71, San Francisco, CA, 1993. Morgan Kaufmann.
- 5 Isabelle de Zegher-Geets. Idefix: Intelligent summarization of a time-oriented medical database. Technical Report KSL-88-34, Knowledge Systems, AI Laboratory, 1987.
- 6 L.M.I. Fugan. *VM2 representing time-dependent relations in a medical setting*. PhD thesis, Stanford University, 1980.
- 7 D. Harmanec, TY Leong, S. Sundaresh, KL Poh, TT Yeo, I. Ng, and TW Lew. Decision analytic approach to severe head injury management. In *Proceedings of the AMIA Symposium*, page 271. American Medical Informatics Association, 1999.
- 8 L. B. Sheiner M. G. Kahn, L. M. Fagan. Combining physiologic models and symbolic methods to interpret time-varying patient data. *Methods of Information in Medicine*, 30(3):167–168, August 1991.
- 9 D. Hadjipanayi M.A. Karaolis, J.A. Moutiris and C.S. Pattichis. Assessment of the risk factors of coronary heart events based on data mining with decision trees. *Information Technology in Biomedicine, IEEE Transactions on*, 14(3):559–566, 2010.
- 10 T.K. Moon. The expectation-maximization algorithm. *Signal Processing Magazine, IEEE*, 13(6):47–60, 1996.
- 11 K. Murphy et al. The bayes net toolbox for matlab. *Computing science and statistics*, 33(2):1024–1034, 2001.
- 12 Michael Ramati and Yuval Shahar. Probabilistic abstraction of multiple longitudinal electronic medical records. In *Proceedings of the 10th conference on Artificial Intelligence in Medicine, AIME'05*, pages 43–47, Berlin, Heidelberg, 2005. Springer-Verlag.
- 13 C. Rose, C. Smaili, and F. Charpillat. A dynamic bayesian network for handling uncertainty in a decision support system adapted to the monitoring of patients treated by hemodialysis. In *Tools with Artificial Intelligence, 2005. ICTAI 05. 17th IEEE International Conference on*, pages 5–pp. IEEE, 2005.
- 14 A. Salatian and J. Hunter. Deriving trends in historical and real-time continuously sampled medical data. *Journal of intelligent information systems*, 13(1):47–71, 1999.
- 15 Y. Shahar and M.A. Musen. Knowledge-based temporal abstraction in clinical domains. *Artificial intelligence in medicine*, 8(3):267–298, 1996.
- 16 M. Stacey and C. McGregor. Temporal abstraction in intelligent clinical data analysis: A survey. *Artificial Intelligence in Medicine*, 39(1):1–24, 2007.
- 17 M.A.J. Van Gerven, B.G. Taal, and P.J.F. Lucas. Dynamic bayesian networks as prognostic models for clinical patient management. *Journal of biomedical informatics*, 41(4):515–529, 2008.
- 18 Yanping Xiang and Kim-Leng Poh. Time-critical dynamic decision making. In *UAI*, pages 688–695, 1999.