Predicting with Variables Constructed from Temporal Sequences

Mehmet Kayaalp
Center for Biomedical Informatics
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA 15213
kayaalp@acm.org

Gregory F. Cooper
Center for Biomedical Informatics
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA 15213
gfc@cbmi.upmc.edu

Gilles Clermont
Department of Anesthesiology
School of Medicine
University of Pittsburgh
Pittsburgh, PA 15213
clermontg@anes.upmc.edu

Abstract

In this study, we applied the local learning paradigm and conditional independence assumptions to control the rapid growth of the dimensionality introduced by multivariate time series. We also combined various univariate time series with different stationary assumptions in temporal models. These techniques are applied to learn simple Bayesian networks from temporal data and to predict survival probabilities of ICU patients on every day of their ICU stay.

1 INTRODUCTION

Temporal modeling is important for a variety of domains ranging from physical sciences to market analysis. For problems that are intrinsically temporal, one needs a robust methodology to provide consistent and reliable temporal decision support.

This paper addresses two key questions in stochastic process modeling: (**1**) How can the rapid growth of the dimensionality introduced by multivariate time series be controlled? (**2**) How can models with various stationarity assumptions be combined?

The methodology developed and evaluated in this study was based on one clinical question: What is an intensive care unit (ICU) patient's chance of survival over the next few days, given all of his/her available temporal measurements that have indicated the physiologic condition of the patient? More specifically, the task is to predict probabilities $(P_1, P_2, ..., P_6)$ of survival of a given patient during the following six mutually exclusive temporal intervals, respectively: 0–1, 1–3, 3–7, 7–15, 15–31, and 31–63 days in the future, where 0 denotes the current day. These clinical predictions may be of interest to a physician at the end of each day of ICU stay of the patient.

In this study, we used a database of physiologic and outcome variables collected on 1,449 patients admitted to 40 different ICUs in May 1995. The database contains 11,418 records, *i.e.*, on average 7.9 records per patient. The temporal granularity of variables is fixed at one day since each record contains one day of collected data on one patient. The data were originally collected for a prospective study to evaluate a newly established Sequential Organ Failure Assessment (SOFA) score that has been used to assess the incidence and severity of organ dysfunction or failure of ICU patients (Vincent and others 1998).

The database contains 25 temporal variables (see Table 1). The original dataset also contains atemporal data, which we did not use in this study, so that we can focus on temporal sequences and ensure that changes in prediction performance are solely due to the newly constructed variables (which we will call patterns) as proposed in the presented methodology.

We discretized patient variables that were continuous in the database based on medical knowledge and their statistical variances observed in the sample population. The third author of this report filled in missing SOFA system values by extrapolating the existing values of the patient variables based on his medical knowledge and judgment. Eighteen percent of values of all other temporal variables were still missing, to which we assigned a separate categorical value, *unknown*. Data collection was limited to 33 days of ICU stay, since only 9 of 1,449 patients stayed in the ICU for more than 33 days.

We define a *patient case* as the physiologic state of a patient on a given day, considering all available temporal data collected during ICU stay of the patient up to and including that given day. For example, a patient in the ICU on day d has cases $(\mathfrak{C}_1, \mathfrak{C}_2, ..., \mathfrak{C}_d)$, where $\mathfrak{C}_{i+1}$ subsumes $\mathfrak{C}_i$, and $i = 1, 2, ..., d-1$. We divided the entire dataset into 65 percent for training (7,388 cases on 949 patients), leaving 35 percent for testing (4,030 cases on 500 patients). We developed patient-specific simple

Bayes models that are learned separately for each patient case using the statistics of training cases (records). We used the area under the receiver operating characteristics (ROC) curve to assess model performance.

Table 1: Temporal variables of the SOFA patient database. Arities of variables are presented in the third column. Arity indicates the number of different values that each discrete variable can take.

Temporal Variable	Arity	Acronym
Oxygenation index	4	pO_2/fiO_2
Mechanical ventilation	2	rsup
Platelet count	4	plat
Bilirubin	3	bili
Mean arterial pressure	4	pam
Dopamine dosage	3	dopa
Dobutamine dosage	3	dobuta
Epinephrine dosage	3	epin
Norepinephrine dosage	3	norepi
Glasgow coma scale	4	gcs
Blood urea nitrogen	5	urea
Serum creatinine	5	creat
Urine output	4	urin
White blood cell count	4	wbc
Heart rate	4	hr
Temperature	4	temp
Sepsis related surgery	2	su
Presence of infection	2	infect
SOFA neurological	6	sofaneuro
SOFA respiratory	6	sofapulm
SOFA cardiovascular	6	sofacard
SOFA hematological	6	sofacoag
SOFA hepatic	6	sofaliver
SOFA renal	6	sofarenal
SOFA total	6	sofatotal

2 BACKGROUND

In an earlier study using the same database along with 8 atemporal variables, we predicted patient mortality at ICU discharge by creating nonstationary and stationary models (Kayaalp, Cooper, and Clermont 2000). The model-building process was based on the standard supervised-learning paradigm, *i.e.*, learning a global model from a training set, where we used a Bayesian scoring metric as defined in (Cooper and Herskovits 1992).

In this study, we used the local learning paradigm. Local learning (*a.k.a.* lazy or instance-based learning) methods let us induce a model using the available data of the test case in question. Although parameters are learned from the training data, the model is optimized specifically to predict the test case in question. Since the target model is an approximation on the local (test) data, those methods are called local learning algorithms.

A stochastic process is defined as (strongly) *stationary* if the probability density functions generated by this stochastic process are the same for all temporal sequences $(t_{i+1}, t_{i+2}, ..., t_{i+n})$, where $i \geq 0$ and $n > 0$ (Jenkins and Watts 1968). For a stationary univariate time series of length $n > 0$, Equation (1) holds for all $i \geq 0$ and any temporal displacement constant $k \geq 0$.

$$P(x_{i+1}, x_{i+2}, ..., x_{i+n}) = P(x_{i+1+k}, x_{i+2+k}, ..., x_{i+n+k}) \quad (1)$$

In this paper, we represent the values of any temporal variable with a lower case letter and a subscripted integer denoting the time stamp of the variable value. For example, $P(X(t) = x_t, X(t+1) = x_{t+1})$ will be abbreviated as $P(x_t, x_{t+1})$, and these expressions denote the joint probability of two successive values of variable X at times t and $t+1$.

A stationary univariate time series model $\mathcal{M}$ with a sequence of $i+1$ data points assumes that x_t is a stochastic function of the sequence $(x_{t-1}, x_{t-2}, ..., x_{t-i})$, and it is conditionally independent of any other factors, given the sequence $(x_{t-1}, x_{t-2}, ..., x_{t-i})$ and the model $\mathcal{M}$; *i.e.*, $P(x_t \mid x_{t-1}, x_{t-2}, ..., x_{t-i}, \mathcal{M}) = P(x_t \mid x_{t-1}, x_{t-2}, ..., x_{t-k}, \mathcal{M})$, where $k > i$. In this report, the term "stationarity assumption" refers to this conditional independence assumption, given a sequence of i successive data points. A Markov chain is a special case of this class of models, where $i = 1$.

Our earlier study showed that nonstationary models perform quite well if the applicable sample size is large enough (Kayaalp, Cooper, and Clermont 2000). However, as time series get longer, the predictive performances of nonstationary models decrease rapidly, due to the exponentially increasing parameter space.

In the current study, a set of new binary variables was constructed from each unique, univariate time series of a length between 1 and 33 time points. Our approach can be considered as a type of constructive induction, creating new variables from existing ones (Pazzani 1996; Bloedorn and Michalski 1998). It can also be seen as a sequence processing and matching technique, which has been used in a variety of domains including information theory (Shannon 1948), bioinformatics (Searls 1993), speech recognition, and text processing (Nevill-Manning 1996). Various methods for representing different stationarity assumptions in the context of short-term memory[1] have also been studied in research on machine learning (Ron, Singer, and Tishby 1996), and recurrent

[1] A memory model is a stochastic function defined by past events. It determines the number of data points to be stored, the resolutions of those data points and their dependence relations.

neural networks (Mozer 1993), among others. But in this study we go one step further and use sequences of various lengths in the same model, combining different stationarity assumptions.

3 METHODS

One key issue in prediction problems with high dimensionality (as in multivariate time series analysis) is representation. The approach presented below reduces the parameter space by (1) representing univariate time series with simpler variables, (2) applying the local learning paradigm, and (3) using conditional independence assumptions.

A discrete multivariate parameter space is determined by the number of variables and their arities. The number of parameters in this parameter space is equal to the number of joint probabilities. For time series models, the time dimension must be taken into account as well. In our database, we have four binary (including the outcome variable of interest), five ternary, eight 4-ary, two 5-ary, and seven 6-ary variables (see Table 1), which translates to $2^4 3^5 4^8 5^2 6^7 \cong 2^{51} \cong 10^{15}$ possible *atemporal* variable-value combinations, which is the size of the atemporal parameter space, when no independence is assumed. The size of the parameter space of a stationary time series with a fixed sequence length d is 2^{51d} without assuming any independence.

Our first reduction of the parameter space comes with a constructive induction approach using the local learning paradigm: Instead of building a single global model and applying it to *all* test cases uniformly, we induced a separate, *local* model for each patient case; the learning process can therefore be called patient-specific. We built new variables from *univariate* time series observed in each patient case. The newly constructed variables are called "patterns." In this report, a pattern is defined as a list of equidistant temporal values of a variable. For example, the body temperature of a patient who stayed in an ICU for three days may have the temperature pattern $\mathcal{P}_{temp1}$ = (*high*, *high*, *normal*). When the list contains a single temporal value, we call it an "elementary pattern," which corresponds to a regular, time-stamped variable.

In this study, we evaluate each pattern $\mathcal{P}$ as a binary variable; in a given data stream, it is either present or not. For example, patterns (*high*), (*normal*), (*high*, *high*), (*high*, *normal*) and (*high*, *high*, *normal*) are positive for the above ICU patient example in the previous paragraph, whereas patterns (*low*), (*normal*, *high*), and (*high*, *high*, *normal*, *normal*) are negative, since they are not observed in the patient data.

If body temperature is the only patient variable and we need to predict the chance of survival of a patient with the temperature pattern $\mathcal{P}_{temp1}$, we should compute $P(C \mid \mathcal{P}_{temp1})$, where C denotes the survival of the patient. Since the length of the value sequence in $\mathcal{P}_{temp1}$ is provided with the patient in question, the probability of observing the exact pattern $P(\mathcal{P}_{temp1})$ may be estimated on the relevant sample of temperature patterns, which is the set of temperature value sequences of the same length as $\mathcal{P}_{temp1}$, *i.e.*, {(*low*, *low*, *low*), (*low*, *low*, *normal*), (*low*, *low*, *high*),…, (*high*, *high*, *high*)}.

In this study, the length of the sequence in a pattern is called aggregation level and denoted as $agg(\mathcal{P})$. Given $\mathcal{P}_i = (x_1, x_2, ..., x_n)$, $agg(\mathcal{P}_i) = n$. In the above example, $agg(\mathcal{P}_{temp1}) = 3$.

The frequency statistic of a pattern $\mathcal{P}_i$ with the aggregation level $agg(\mathcal{P}_i)$ is collected from the sample of patterns $\{\mathcal{P}_j\}$ of the *same* variable with the *same* level of aggregation; *i.e.*, $\mathcal{P}_i \in \{\mathcal{P}_j \mid j = 1,2,...,J \quad \wedge agg(\mathcal{P}_j) = k\}$, where k is constant, and J is the number of patterns in $\{\mathcal{P}_j\}$. If it is a univariate temporal pattern, where the arity of variable is a, then $J = a^k$. In the temperature example, the cardinality of the pattern set, to which $\mathcal{P}_{temp1}$ belongs, is 3^3.

The probability of $\mathcal{P}_i$ in an arbitrary univariate sequence of length $agg(\mathcal{P}_i)$ can be estimated as

$$P(\mathcal{P}_i) = \frac{n(\mathcal{P}_i)}{\sum_{j=1}^{J} n(\mathcal{P}_j)} \tag{2}$$

where $n(\cdot)$ returns the frequency count of its attribute. For example, if $\mathcal{P}_{temp1}$ was observed in 10 patient cases and all other temperature patterns with the same aggregation level were observed in 90 patient cases, then $P(\mathcal{P}_{temp1}) = 0.1$.

Joint probability of the pattern $\mathcal{P}_i$ and outcome variable C can be estimated as

$$P(C, \mathcal{P}_i) = \frac{n(C, \mathcal{P}_i)}{\sum_{j=1}^{J} n(C, \mathcal{P}_j)} \tag{3}$$

Using Equations (2) and (3), we can compute conditional outcome probability.

$$P(C \mid \mathcal{P}_i) = \frac{P(C, \mathcal{P}_i)}{P(\mathcal{P}_i)} \tag{4}$$

A database of patterns is built from training patient cases. Recall that a patient case on day d contains all data of d records; hence, it has d consecutive daily values

$(x_1, x_2, ..., x_d)$ measured for each variable. There are d patterns $\{(x_d), (x_{d-1}, x_d), ..., (x_1, x_2, ..., x_d)\}$ for each variable associated with this patient case. The *pattern set* of a patient case with v variables and d days of history consists of $v \times d$ patterns. By this definition, all sequences that do not include the last day's measurement x_d are excluded from the pattern set. The pattern set along with frequency statistics of all patterns in the training data constitutes the pattern database that we use to construct patient-specific models.

Since all patterns are binary, the size of a parameter space that is specific to a patient case with v regular temporal variables and d days is 2^{vd}. In addition to 25 temporal variables in our database, there is one binary response variable (mortality) in each model; thus, the size of a patient-specific parameter space is equal to 2^{25d+1}. This is approximately 2^{25} times reduction of the parameter space; recall that the size of the parameter space is 2^{51d} when data are represented as a stationary multivariate time series of length d without assuming independence.

Our second reduction of the parameter space comes with the conditional independence assumption: When patterns are assumed to be conditionally independent, given the binary outcome variable of interest, the size of the exponential parameter space is reduced to a polynomial $2^2 vd$. For the current database, this number is $100d$. Notice that, conditional independence is assumed between patterns, not between the events[2] in a pattern.

The resulting temporal model is a simple Bayes model: $P(C \mid \mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_m) \propto P(C) \prod_{i=1}^{m} P(\mathcal{P}_i \mid C)$, where each $\mathcal{P}_i$ denotes a pattern observed in a given patient case and is included in the model, and C represents the outcome variable of interest. Although it is a violation of conditional independence assumption, in our experiments we did not restrict models to the set of patterns that are mutually exclusive.

Given a database of patterns, model selection is reduced to a pattern selection (variable selection) process in a simple Bayes modeling approach. The following steps summarize the pattern selection process that we performed in this study:

1. All patterns in a given test patient case were identified.
2. The probability of each pattern was estimated using the frequency statistics that were collected from training patient cases and represented in the pattern database.
3. Each pattern along with the outcome variable of interest was evaluated separately for its predictive significance using the area under the ROC curve, which is a measure of the prediction performance of a model.
4. Patterns whose outcome prediction performances yielded ROC areas smaller than 50 percent were eliminated.
5. Patterns were rank-ordered and m patterns with the highest ROC scores were selected for inclusion into the final model, where m is determined by a simple validation process discussed below.

Using a small validation set[3] of 330 patient cases, we searched for m, an optimal number for patterns to include in simple Bayes models.

As described in the Introduction, the models were built to predict $(P_1, P_2, ..., P_6)$ the survival chance of each ICU patient at six mutually exclusive temporal intervals of their ICU stay. Our preliminary results as evaluated in the next section indicate that m the optimal size of the pattern set used in these models is equal to 128. m is an upper bound only; obviously, all models could not have 128 patterns, since the number of patterns in each patient case can be at maximum $25d$, where d is the number of days in the ICU. m can be less than $25d$, because patterns whose outcome predictive performances yielded ROC areas less than 50 percent were excluded from the pattern set during the validation process.

Predictions $(P_1, P_2, ..., P_6)$ of the final models of each patient case were also evaluated with the same ROC metric.

The results were produced on three parallel running processes on three 600 MHz Intel Pentium II based Linux machines in approximately one day. The experiment required 93 MB system memory.

4 RESULTS

Our preliminary results indicate that patient-specific models with a maximum of 128 patterns perform best, yielding areas under the ROC curves between 75 and 80 percent for all 6 predictions (see Figure 1).

A single-pattern model evaluated in this study is a bivariate Bayesian network, in which the outcome variable of interest is dependent on one pattern. We compared multi-pattern models with single pattern

[2] An event is an observation that is measured at a specific time point and represented as a variable value in a time series.

[3] Validation and test sets are mutually exclusive. Patient cases in the validation set are randomly selected from the training set.

models, since the latter is the best representative of a bivariate temporal Bayesian network with regular temporal variables, and all models that were found in our earlier study most predictive of survival of ICU patients (Kayaalp, Cooper, and Clermont 2000) were also bivariate temporal Bayesian networks with regular temporal variables. Recall that a regular temporal variable is equivalent to an elementary pattern of that variable.

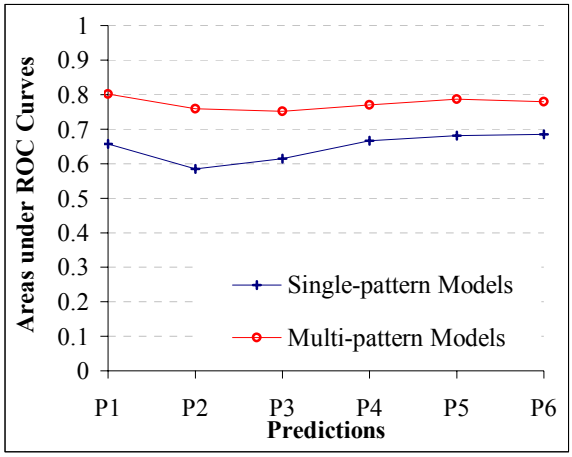

Figure 1: Prediction Performances of Single-Pattern *vs.* Multi-Pattern Models.

Table 2 shows percentage of patterns, each of which was found the most significant in a set of predictions. The patterns shown in Table 2 cover 85 percent of all patterns used in single-pattern models. Recall that the last value in every pattern corresponds to a data point observed on the last day of the patient case in question. Each pattern in Table 2 is presented with the variable name of the pattern and a sequence of variable values $(...,x_{d-1},x_d)$, where x_d is the data point observed on the last day. Therefore, "rsup (2,2)" refers to the use of mechanical ventilation during the last two days of ICU stay, value "(1)" associated with SOFA patterns, indicates that the functional parameters of the associated organ systems are within physiological ranges, and "urin (2)" indicates low urine output and renal system dysfunction.

Table 2 shows that use of mechanical ventilation is a dominant predictor of the patient survival between days 2 and 63 following the day when the prediction is made. Cardiac system related SOFA score is the most dominant on the first day and during the second half of the first month following the prediction day. Renal system related patterns are significant during the first day and the second month following the prediction day.

Table 2: Percentage of patterns found most significant, with largest ROC areas in predictions P_1 through P_6. The first column contains patterns. Numbers in parentheses are data sequences that appeared in those patterns.

	P_1	P_2	P_3	P_4	P_5	P_6
rsup (2,2)	3	46	47	0	25	37
rsup (1)	1	32	32	7	7	8
rsup (2)	0	10	10	30	5	6
sofacard (1)	50	0	0	55	55	0
urin (2)	11	0	0	0	0	0
sofarenal (1,1)	6	0	0	0	0	29
others	29	13	11	9	9	20

We built 22,152 multi-pattern models for 3,692 patient test cases by using 8,469 unique patterns. Although only 18 percent of patterns were uniform sequences such as $(2,2,...,2)$, 91 percent of the time only uniform patterns were selected into the models. We were expecting that predictive patterns would capture worsening conditions of decompensating patients, but, instead, patterns indicating stability were selected the most. One reason why we could not observe many patterns of change may have been due to the scoring function that we set in our pattern selection process. The current scoring function maximizes the area under the ROC curve, which is a function of the sensitivity and specificity of the model predictions. In the training database, the survival rates of patients decrease slowly, from 0.97 to 0.73, while the prediction range gets longer. It might be possible to capture patterns of decompensating patients by changing the scoring function.

5 CONCLUSIONS

In this study, we addressed two key issues:

(1) Clinical prediction problems represented in multivariate time series are subject to the curse of dimensionality. The local learning paradigm along with the constructive induction approach and conditional independence assumptions, can reduce the global parameter space to a local, smaller parameter space given the data of a single patient. Instead of considering all combinations of possible time series, we constructed a new set of variables only from those patterns that appeared in the patient case in question.

(2) How can time series with various stationarity assumptions be combined? By constructing patterns from time series with various lengths, hence with different stationarity assumptions, and building models using those patterns, we could represent and combine different dependence relationships observed in univariate event sequences.

In this preliminary study, we limited the focus of research to the above stated two points, and tried not to include any additional degree of freedom, such as search on multivariate pattern space and search on unrestricted space of Bayesian network structures. When such searches are performed effectively, more expressive, predictive patterns, and better model structures are likely to be found; however, predictive results of the presented method with such extensions would then be strongly affected by the degree of the effectiveness of the heuristics used in those additional search procedures.

6 FUTURE STUDIES

In this study, we used only an aggregation technique to construct variables from time series patterns. We are planning to use some abstraction techniques to combine patterns that are similar in nature. Abstraction techniques would enable us not only to utilize the available sample population more effectively but also to include other combinations of time series that we excluded in the presented study, without any additional burden of computational complexity.

We also plan to extend our approach to use temporal multivariate patterns in hierarchical models and apply prequential analysis (Dawid 1984).

Acknowledgements

We thank Drs. Jean-Louis Vincent, Rui Moreno, and the European Society of Intensive Care Medicine for the provision of the SOFA dataset and their support of this study. We also thank to our anonymous reviewers for their constructive questions and remarks.

This work was supported by the National Library of Medicine with the grant "Integrated Advanced Information Management Systems" No. G08-LM06625. Research support for Greg Cooper was provided in part also by grants Nos. R01-LM06696 and R01-06759 from the National Library of Medicine and by grant No. IIS-9812021 from the National Science Foundation.

References

Bloedorn, E. and Michalski, R.S., 1998, Data-Driven Constructive Induction: IEEE Intelligent Systems, 13, p. 30–37.

Cooper, G.F. and Herskovits, E., 1992, A Bayesian Method for the Induction of Probabilistic Networks from Data: Machine Learning, 9, p. 309–347.

Dawid, P.A., 1984, Present Position and Potential Developments: Some Personal Views. Statistical Theory. The Prequential Approach: Journal of Royal Statistical Society A, 147, p. 278–292.

Jenkins, G.M. and Watts, D.G., 1968, Spectral Analysis and Its Applications. San Francisco, CA, Holden-Day,

Kayaalp, M.M, Cooper, G.F., and Clermont, G., 2000, Predicting ICU Mortality: A Comparison of Stationary and Nonstationary Temporal Models. Proc. AMIA 2000 Symposium, p. 418–422. Los Angeles, CA.

Mozer, M.C., 1993, Neural Net Architectures for Temporal Sequence Processing: Weigend, A. S. and Gershenfeld, N. A. Time Series Prediction: Forecasting the Future and Understanding the Past. Addison–Wesley.

Nevill-Manning, C.G., 1996, Inferring Sequential Structure: University of Waikato.

Pazzani, M.J., 1996, Constructive Induction of Cartesian Product Attributes. Information, Statistics and Induction in Science, Melbourne.

Ron, D., Singer, Y., and Tishby, N., 1996, The Power of Amnesia: Learning Probabilistic Automata with Variable Memory Length: Machine Learning, 25, p. 117–149.

Searls, D.B., 1993, The Computational Linguistics of Biological Sequences: Hunter, L., p. 47–120. Artificial Intelligence and Molecular Biology. MIT Press: Cambridge, MA.

Shannon, C.E., 1948, A Mathematical Theory of Communication: The Bell System Technical Journal, 27, p. 379–423, 623–656.

Vincent, J.-L.M.P.F., de Mendonca, A.M., Cantraine, F.M., Moreno, R.M., and Blecher, S.M., 1998, Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: Results of a multicenter, prospective study: Critical Care Medicine, 26, p. 1793–1800