Predicting with Variables Constructed from Univariate Temporal Sequences

Mehmet Kayaalp, Gregory F. Cooper, and Gilles Clermont
University of Pittsburgh
kayaalp@acm.org, gfc@cbmi.upmc.edu, clermontg@anes.upmc.edu

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 on any of the following six mutually exclusive temporal intervals: 0-1, 1-3, 3-7, 7-15, 15-31, and 31-63 days.

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. Each record contains one day of collected data on one patient; *i.e.*, the temporal granularity of variables is fixed at one day. 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 [7].

The database contains 25 temporal variables (see Table 1). We discretized patient variables that were continuous in the database based on medical knowledge and their statistical variances observed in the sample population. Data collection was limited to 33 days of ICU stay.

We define a *patient case* as the physiologic state of a patient on a given day of his/her stay in the ICU, considering all of his/her available temporal data collected during his/her ICU stay up to and including that given day. For example, a patient in the ICU on day d has cases $(D_1, D_2, ..., D_d)$, where D_{i+1} subsumes D_i, and $i = 1, 2, ..., d-1$. In database terms, D_1 is represented in the record of the first day, D_2 in the records of the first and second days, and so on. We developed patient-specific simple Bayes models that are learned separately for each patient case using the statistics of 7,388 training cases (records) of 949 patients. We used the area under the receiver operating characteristics (ROC) curve to assess model performance.

No	Temporal Variable	Arity & Acronym	
1	Oxygenation index	4	pO_2/fiO_2
2	Mechanical ventilation	2	rsup
3	Platelet count	4	plat
4	Bilirubin	3	bili
5	Mean arterial pressure	4	pam
6	Dopamine dosage	3	dopa
7	dobutamine dosage	3	dobuta
8	Epinephrine dosage	3	epin
9	Norepinephrine dosage	3	norepi
10	Glasgow coma scale	4	gcs
11	Blood urea nitrogen	5	urea
12	Serum creatinine	5	creat
13	Urine output	4	urin
14	White blood cell count	4	wbc
15	Heart rate	4	hr
16	Temperature	4	temp
17	Sepsis related surgery	2	su
18	Curr. state of infection	2	infect
19	SOFA neurological	6	sofaneuro
20	SOFA respiratory	6	sofapulm
21	SOFA cardiovascular	6	sofacard
22	SOFA hematological	6	sofacoag
23	SOFA hepatic	6	sofaliver
24	SOFA renal	6	sofarenal
25	SOFA total	6	sofatotal

Table 1 Temporal variables of the SOFA patient database. Values in the third column indicate arities of variables, *i.e.*, the number of different values that each discrete variable can take.

BACKGROUND

In an earlier study using the same database, we predicted patient mortality at ICU discharge by creating non-stationary and stationary models [4]. 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

model scoring metric as defined by Cooper and Herskovits [2]. Unlike the approach of the present study, the learning process in our earlier study was global; *i.e.*, a single model was built for each stationarity assumption and applied to all test cases.

A stochastic process is defined as (strongly) stationary if it is time invariant with respect to any arbitrary time shift t in either direction, *i.e.*, $P(x_1, x_2, ..., x_i) = P(x_{1+t}, x_{2+t}, ..., x_{i+t})$, where i determines the number of data points that are dependent. A stationary univariate time series model M with a sequence of i data points assumes that x_t is a stochastic function of the sequence $(x_{t-1}, x_{t-2}, ..., x_{t-i})$, *i.e.*, x_t is conditionally independent of any other factors, given the sequence $(x_{t-1}, x_{t-2}, ..., x_{t-i})$ and the model M. 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 non-stationary models perform quite well if the applicable sample size is large enough. However, as time-series get longer, the predictive performances of non-stationary models decrease rapidly, due to the exponentially increasing model dimensionality.

In the current study, a set of new binary variables was constructed from each unique, univariate time series with the number of data points ranging from 1 to 33. Our approach can be considered as a type of constructive induction, creating new variables from existing ones [1,6]. It can also be seen as a sequence matching technique, which has been used in a variety of domains including coding theory, bioinformatics, speech recognition, text processing, and compiler optimization. Various methods for representing different stationarity assumptions in the context of short-term memory[1] have also been studied in research on recurrent neural networks [5].

METHODS

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

[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.

A discrete multivariate model space is determined by the number of variables and their arities, *i.e.*, the number of different values that each variable can take. 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 model space. The size of the model space of a stationary time-series with a fixed sequence length d is 2^{51d}.

Our first reduction of the model 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 data points measured successively at equal temporal intervals. When the list contains a single data point, we call it an "elementary pattern," which corresponds to a regular, time-stamped variable.

Each pattern T is a binary variable, and it is either present or absent for any given patient case. The aggregation level of T, denoted as $agg(T)$, is equal to the length of a pattern sequence T; *i.e.*, $agg(T) = n$, where $T = (x_1, x_2, ..., x_n)$ The statistic of a pattern T_i with the aggregation level $agg(T_i)$ is collected from the sample of patterns $\{T_j\}$ of the *same* variable with the *same* level of aggregation; *i.e.*, $T_i \in \{T_j \mid j = 1, 2, ..., J \wedge agg(T_j) = k\}$, where k is constant, and J is the number of patterns in $\{T_j\}$. So, the probability of the occurrence of T_i in such a sample can be estimated as $n(T_i) \Big/ \sum_{j=1}^{J} n(T_j)$, where $n(\cdot)$ returns the frequency count of its attributes.

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 values $(x_1, x_2, ..., x_d)$ measured for each variable.[2] 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. A patient case with v variables has $v \times d$ patterns on day d. By this

[2] For every missing data point, we assign a categorical value indicating its missing status.

definition, all sequences that do not include the last day's measurement (*i.e.*, x_d) are excluded from the pattern set. Since all patterns are binary, the dimension of a model 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 model space is equal to 2^{25d+1}.

Our second reduction of the model 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 model 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 in a pattern.

The resulting temporal model is a simple Bayes model: $P(C \mid T_1, T_2, ..., T_m) \propto P(C) \prod_{i=1}^{m} P(T_i \mid C)$, where each T_i denotes a pattern observed in a given patient case and is included in the model, and C represents the outcome variable of interest.

Given a database of patterns, the model selection is reduced to a variable selection process in a simple Bayes modeling approach. First, all patterns in a given test patient case were identified. The statistic of each pattern was collected from training patient cases. Each pattern along with the outcome variable of interest was used to create a bivariate model. Each bivariate model was evaluated using the area under the ROC curve, which is a function of the prediction performance of the model. Patterns of models whose prediction performances yielded ROC areas smaller than 50 percent were eliminated. Remaining 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. Predictions of the final models of each patient case were also evaluated with the same ROC metric.

RESULTS AND CONCLUSIONS

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. This validation process was rather crude and serves only to provide preliminary results. Results to be reported in the final report will be obtained through a more extensive cross-validation process.

[3] Validation and test sets are mutually exclusive.

As described in the Introduction of the present report, the models were built to predict the survival chance of each ICU patient at six mutually exclusive temporal intervals of their ICU stay. 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 six predictions (see Figure 1). The size of the pattern set used in these models indicates an upper bound only; obviously, all models could not have 128 patterns, since the number of patterns in each patient case can be $25d$ at maximum, where d is the number of days in the ICU, and patterns whose bivariate models yielded ROC areas less than 50 percent were also excluded during the validation process.

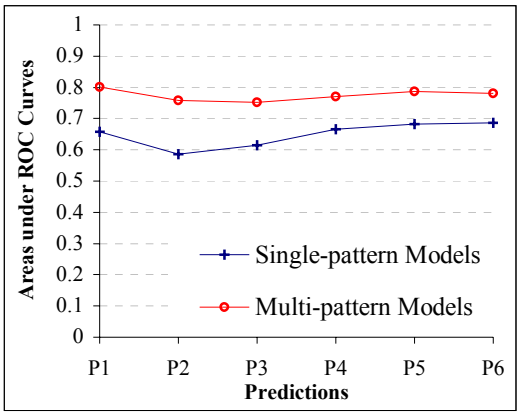

Figure 1 Prediction performances of single-pattern *vs.* multi-pattern models.

We compared multi-pattern models with single-pattern models, which correspond to the temporal models found most predictive in determining ICU mortality in our earlier study [4].

Table 2 shows frequencies of patterns, each pattern 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

	P1	P2	P3	P4	P5	P6
rsup (2,2)	101	1685	1721	0	909	567
rsup (1)	20	1168	1193	244	241	121
rsup (2)	17	354	356	1101	178	91
sofacard (1)	1845	0	0	2030	2044	0
urin (2)	395	0	0	0	0	0
sofarenal (1,1)	238	0	0	0	0	450
others	1077	486	423	318	321	314

Table 2 Frequencies of patterns found most significant, with largest ROC areas in predictions P1 through P6. The first column contains patterns. Numbers in parentheses are data sequences that appeared in those patterns.

to a data point observed on the last day of the patient case in question. Each pattern in Table 2 is presented with a variable name of the pattern and a sequence of data points $(..., x_{d-1}, x_d)$, where x_d is the data point observed on the last day. Therefore, "**rsup (2,2)**" refers to the usage of mechanical ventilation (see Table 1) during the last two days of ICU stay. In Table 2, value 1 in parentheses associated with SOFA patterns indicates that the functional parameters of the associated organ systems are within physiological ranges. "**urin (2)**" indicates low urine output, *i.e.*, renal system dysfunction.

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 the patterns that were selected to be used in the models were surprisingly uniform. We were expecting that predictive patterns would capture worsening conditions of decompensating patients, but, instead, patterns indicating stability were selected most. One reason why we could not observe many patterns of change may have been due to the utility function that we set in our pattern selection process. The current utility function maximizes the area under the ROC curve, which is a linear sum of the sensitivity and specificity of the model predictions. In the training database, the survival rates of patient cases decrease slowly, from 0.97 to 0.73, while the prediction range gets longer. Had we set our utility function to capture mortality with higher sensitivity, the resulting patterns might have been more specific to the patient cases that were decompensating. In the final version of this paper, a detailed analysis will be included.

The test 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.

In this study, we addressed two key issues: Clinical problems represented in multivariate time-series are subject to the curse of dimensionality. The local learning paradigm along with constructive induction approach and conditional independence assumption can reduce the global model space to a smaller model space of a single patient's data. 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.

The other key issue addressed in this study was how could 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 relations observed in univariate event sequences.

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 broaden the model space by including additional sets of time-series without increasing computational complexity. We also plan to use temporal patterns in hierarchical models and search for multivariate interactions. Our research may also benefit by incorporating prequential priors while each patient case evolves [3]. We plan to include a prequential analysis in the final version of this paper.

REFERENCES

[1] Bloedorn, E. and Michalski, R.S., Data-Driven Constructive Induction. *IEEE Intelligent Systems*, vol. 13, pp. 30-37, 1998.

[2] Cooper, G.F. and Herskovits, E., A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, vol. 9, pp. 309-347, 1992.

[3] Dawid, P.A. , Present Position and Potential Developments: Some Personal Views. Statistical Theory. The Prequential Approach *Journal of Royal Statistical Society A*, vol. 147, pp. 278-292, 1984.

[4] Kayaalp, M., Cooper, G.F., and Clermont, G., "Predicting ICU Mortality: A Comparison of Stationary and Nonstationary Temporal Models," Proc. *AMIA 2000 Annual Symposium* (in press).

[5] Mozer, M.C. Neural Net Architectures for Temporal Sequence Processing. In: *Time Series Prediction: Forecasting the Future and Understanding the Past*, Eds. Weigend, A.S. and Gershenfeld, N.A. Addison-Wesley, 1993.

[6] Pazzani, M.J., "Constructive Induction of Cartesian Product Attributes," *Information, Statistics, and Induction in Science,* Melbourne, Australia.

[7] Vincent, J.-L.M.P.F., de Mendonca, A.M., Cantraine, F.M., Moreno, R.M., and Blecher, S.M., 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*, vol. 26, pp. 1793-1800, 1998.