CinC Challenge: Cluster Analysis of Multi-Granular Time-series Data for Mortality Rate Prediction

Jianfeng Xu[1,2], Dan Li[2], YuanjianZhang[1], Admir Djulovic[2], Yu Li[1],Youjie Zeng[3,1]

[1]Software School, Nanchang University, China
[2]Department of Computer Science, Eastern Washington University, USA
[3] Department of Application Software Development, Beijing Salien Company, China

Abstract

The goal of this research is to develop novel cluster analysis techniques to identify similarity between ICU time-series data. The results generated by cluster analysis are further used for ICU mortality prediction. To preprocess multi-granular ICU time-series, we proposed a segmentation-based method to divide time-series into several segments. The minimal and maximal values within each segment were captured to maintain the statistical feature of the segment. A weighted Euclidean distance function was in place to evaluate the similarity between two instances and clustering was later used to convert each time-series into a corresponding cluster number. This way, we turned the high dimensional ICU time series data into a 2-dimensional matrix. A rule-based classification model was developed from this 2-dimensional matrix, and the model was used to predict the in-hospital mortality for test cases. The experiments show that above approach is effective in handling ICU time-series data.

1. Introduction

The development of methods for prediction of mortality rates in Intensive Care Unit (ICU) populations has been motivated primarily by the need to compare the efficacy of medications, care guidelines, surgery, and other interventions [1]. The focus of the PhysioNet/CinC Challenge 2012 is to develop methods for patient-specific prediction of in-hospital mortality[1]. Participants will use information collected during the first two days of an ICU stay to predict which patients survive their hospitalizations, and which patients do not.

37 test variables are collected when a patient is in hospital. These 37 variables may be observed once, more than once, or not at all in some cases and each observation has an associated time-stamp indicating the elapsed time of the observation since ICU admission in each case, in hours and minutes. Thus, the project, in nature, can be considered as a problem of multi-dimensional time series data mining in the field of medical information or medical predicting.

The problem of time series pattern recognition has obtained intensive attention in data mining community, due to its prevalence in numerous applications[2-5]. In most applications, especially with the wide-wide, proliferation of multimedia technology, more and more multivariate time series data have appeared[7-14]. These include motion capture data, bio-informatics sequence, financial data, moving object tracking, and many others. Most progress achieved in this area has focused on single granularity time series problems. However, the time-series data collected from ICU has its special features, such as high-dimensionality and multi-granularity, which make the study difficult and challenging. Therefore, there is a need for providing solutions that address this type of time series data. In this paper, we present novel cluster analysis techniques to identify similarity between multi-dimensional and multi-granularity time-series data. The results generated by cluster analysis are further used for ICU mortality prediction.

2. The prediction system

Figure 1 shows the system architecture which consists of three main components, i.e., time series cluster analysis, prediction rule extraction, and mortality rate prediction. First, since the original ICU data include variables collected in different time intervals, the time series cluster analysis component aims at converting the multi granularity data into single granularity, through the statistical analysis of each time series input variable. Second, the prediction rule extraction component uses a traditional classification method to extract decision rules from data set A. Finally, the mortality rate prediction component applies the decision rule to data set B to evaluate the overall performance of the system.

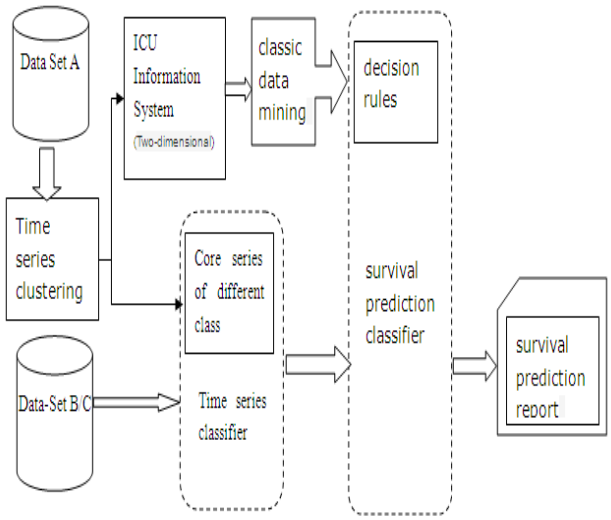

Figure 1. System architecture.

2.1. Similarity computing

Similarity computing is one of the basic research topics in time series data clustering [15-17]. The main issue in similarity computing is that many traditional methods are sensitive to mutual drift in time series data. To address this issue, we implement a segmentation-based method to divide time-series into several segments. The minimal and the maximal values within each segment are captured to maintain the statistical feature of the segment, as shown in Figure 2.

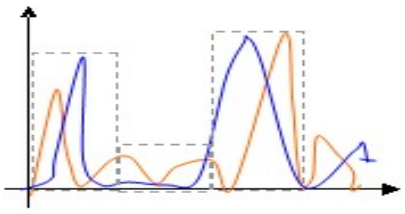

Figure 2. Segmentation of time-series data.

After dividing the original multi-granular time-series into a fixed number of segments, an improved Euclidean distance Function is in place to evaluate the similarity between two ICU time-series instances.

$$D\,d = \frac{\sum_{i=1}^{n}\sqrt{\left(\left(x_{i\max}-y_{i\max}\right)^2+\left(x_{i\min}-y_{i\min}\right)^2\right)}}{\sum_{i=1}^{n}\sqrt{\left(\left(x_{i\max}\right)^2+\left(x_{i\min}\right)^2\right)}}$$

2.2. Clustering algorithm

Based on time slice and time series similarity function, K-means clustering algorithm on multivariate time series data is employed to turn each time series variable input a single cluster number. The centroid of each cluster is recorded, which will later be used on test cases. By turning each input sequence into a single value, we turn the original high-dimensional and multi-granular time series data into a two-dimensional matrix, as shown in table 1 which includes 10 attributes and 23 patient records.

Table 1. Example of 2-dimensional matrix of ICU data

RecordID	HR	NIDias	NISy	RespR	PH	Pa02	PaC02	WBC	Urine	Temp	Death
132539	0	0	0	0	0	0	0	0	0	0	0
132540	0	0	0	0	0	0	0	0	0	0	0
132541	2	2	2	1	2	2	2	2	2	2	0
132543	3	3	3	3	0	0	0	3	3	3	0
132545	4	4	4	4	0	0	0	4	4	4	0
132547	5	5	5	1	5	5	5	5	5	5	0
132548	6	6	6	6	0	0	0	6	6	6	0
132551	7	7	7	1	7	7	7	7	7	7	1
132554	8	8	8	8	0	0	0	8	8	8	0
132555	9	9	9	1	9	9	9	9	9	9	0
132556	5	0	0	4	0	0	0	0	7	0	0
132567	2	7	0	1	0	0	0	2	7	1	0
132568	5	3	3	1	0	0	0	4	4	2	0
132570	3	4	8	0	0	0	0	9	9	8	0
132585	5	9	9	1	1	1	9	2	5	2	0
132588	9	0	0	3	0	0	0	4	7	0	1

2.3. Rule extraction and mortality rate prediction

The two-dimensional array developed in the previous step makes the development of classification model straightforward. As shown in Table 1, HR, NIDIASABP, etc. are input attributes, while Death is the classification attribute. Many traditional classification methods, such as decision trees, IF-THEN rules, support vector machine, etc. can be used to extract decision rules.

Once the classification model is chosen, the mortality rate prediction process is defined as follows:

STEP1: A test record is read from the test dataset which includes 37 time-series sequences.

STEP2: Each time-series sequence is compared with the centroid of each cluster which is derived from the clustering algorithm.

STEP3: Repeat STEP 2 to all 37 input variables and turn 37 time-series sequences into 37 corresponding cluster numbers.

STEP4: Apply the decision rules to the test case to predict the in-hospital death.

3. Experimental results

3.1. Preliminary experiments

Due to time constraint, the preliminary experiments are conducted on 10 input variables recommended by ICU doctors. These variables include PH, NISYSABP, PACO2,PAO2,WBC,TEMP,WBC,URINE,RESPRATE, NIDIASABP. The given ICU Date Set A consisting of 4000 records is divided into three subsets. Each of the subset can be used as training set or test set.

Each time series sequence is first divided into 16 time intervals, and the length of each interval is fixed to three hours. Afterwards, time-series clustering algorithm, rules extraction algorithm, and mortality rates prediction algorithm are performed in proper order.

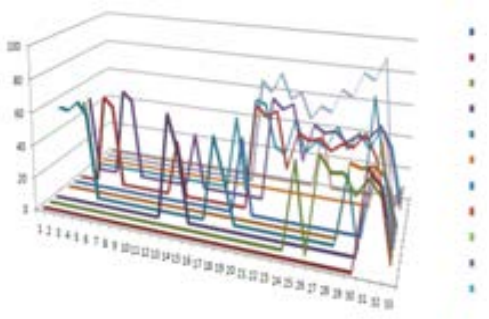

Figure 3-1. Cluster samples 1.

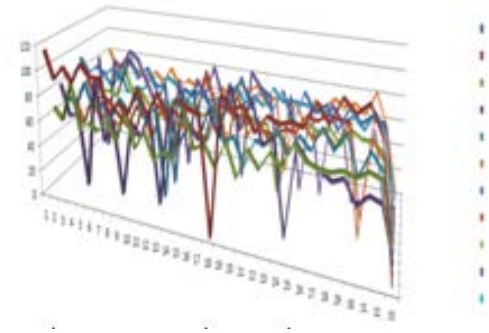

Figure 3-2. Cluster samples 2.

Figure 3-1 and Figure 3-2 show two clusters generated by K-means approach after time-series segmentation of input variable. We only displayed 10 patient instances (i.e., 10 curves) in each cluster to make the picture easy to read.

We can see that, from these two clusters, all the instances within the same cluster shared the similarly features and these features are well-captured by our segmentation-based time-series handling approach.

The preliminary experimental results are as follows: regarding death cases of ICU data set A , the prediction accuracy was 22.77% for Event 1 , and the live prediction accuracy is 75%.

3.2. Second experiments

After Phase 1, we further refine our solutions to improve the prediction accuracy. First of all, even though the 10 variables used in the preliminary testing phase are suggested by domain experts, we cannot simply afford losing the rest 27 given input variables. Thus, in Phase 2, all of the 37 ICU attributes are used improve the experimental results.

Second, since we use fixed number of segments, this requires appropriate missing value handling methods to fill in missing values within certain segments. Different from Phase 1 which we simply used mean values to replace missing values, in Phase 2, we used advanced missing data imputation algorithm to capture the statistical feature in the original data.

Third, X-means clustering algorithms used to replace the original K-means algorithm because it is difficult to determine the value of K and the initial Kcentroids, which normally affect the prediction accuracy.

Similar to Phase 1, the given training dataset A is divided into three subsets with each subset being used once as the test set. Table 2 shows the comparison of our experimental results in Phase 1 and Phase 2.

Table 2. The experimental results.

Test set / results	Experiments 1			Experiments 2		
	Set -a	Set -b	Set -c	Set -a	Set -b	Set -c
death prediction	0.24	0.21	0.22	0.39	0.41	0.43
live prediction	0.72	0.77	0.76	0.86	0.83	0.89

The average results therefore are optimized from 22.77% to 33.08% for death prediction, and from 75% to 86% for live prediction.

Obviously, taking the whole 37 ICU attributes and optimization algorithms into consideration has highly improved the accuracy of the forecasts.

3.3. Further experiments

The further experiments are conducted on the whole 37 input variables. New general descriptor ICUType was added.

The further experimental results are as follows:

regarding death cases of ICU data set A , the average prediction accuracy was 51.525%.

1. ICUType = 1, the number of extraction rules is 118, in a total of 80 deaths, 47 of them were accurately predicted, the accuracy rate was 58.8%;

2. ICUType = 2, the number of extraction rules is 31 , in a total of 43 deaths, 8 of them were accurately predicted, the accuracy rate was 18.6%;

3. ICUType = 3, the number of extraction rules is 115 , in a total of 100 deaths, 63 of them were accurately predicted, the accuracy rate was 63%;

4. ICUType = 4, the number of extraction rules is 109 , in a total of 100 deaths, 66 of them were accurately predicted, the accuracy rate was 66%;

4. Conclusions

Due to the high-dimensionality and multi-granularity features of ICU data, the prediction of mortality rate is a traditional but difficult research topic. This study shows that the segment-based clustering approach is effective in handling ICU time-series data and it is a feasible way for prediction of mortality rates. We plan to further improve our solution by choosing different distance metrics to refine the clustering approach, and developing more effective rule extraction algorithm.

References

[1]http://physionet.org/challenge/2012/ #outcome-relateddescriptors, [Online], Accessed July-03-2012.

[2] GUAN He-Shan, JIANG Qing-Shan, WANG Sheng-Rui Pattern Matching Method Based on Point Distribution for Multivariate Time Series,Journal of Software, Vol.20, No.1,pp.67−79, January 2009,

[3] Singhal A, Seborg DE. Matching patterns from historical data using PCA and distance similarity factors. In: Krogh BH, ed. Proc. of the 2001 American Control Conf. Arlington, 2001,2:1759−1764.

[4] Liu B, Liu J. Multivariate time series prediction via temporal classification. In: Rakesh A, ed. Proc. of the 18th Int'l Conf. on Data Engineering. Washington: IEEE Computer Society, 2002. 268.

[5] Camarinha-Matos LM, Seabra Lopes L, Barata J. Integration and learning in supervision of flexible assembly systems. IEEE Trans. on Robotics and Automation, 1996,12(2):202−219.

[6] Liu HT, Ni ZW, Li JY. An effective algorithm to match similar time series pattern. Journal of Computer-Aided Design & Computer Graphics, 2007,19(16):725−729 (in Chinese with English abstract).

[7] Huang H, Huang K, Hang XS, Xiong FL. Algorithm for fast time-series pattern recovery in a long sequence. Computer Engineering and Applications, 2003,39(21):192−194 (in Chinese with English abstract).

[8] Ge XP, Padhraic S. Deformable Markov model templates for time-series pattern matching. In: Proc. of the 6th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2000. 81−90.

[9] Pratt KB, Fink E. Search for patterns in compressed time series. Int'l Journal of Image and Graphics, 2002,2(1):89−106.

[10] Wang XH. Study on time series similarity and trend prediction [Ph.D. Thesis]. Tianjin: Tianjin University, 2003 (in Chinese with English abstract).

[11] Dong XL, Gu CK, Wang ZG. Research on shape-based time series similarity measure. Journal of Electronics & Information Technology, 2007,29(5):1228−1231 (in Chinese with English abstract).

[12] Wu SC, Wu GF, Wang W, Yu ZC. A time-sequence similarity matching algorithm for seismological relevant zones. Journal of Software, 2006,17(2):185−192 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/185.htm

[13] Kalpakis K, Gada D, Puttagunta V. Distance measures for effective clustering of ARIMA time-series. In: Nick C, ed. Proc. of the IEEE Int'l Conf. on Data Mining. Washington: IEEE Computer Society Press, 2001. 273−280.

[14] Zhang J, Wu SC, Wang W. Research of data mining method on multivariate time series. Computer Engineering and Design, 2006, 27(18):3364−2266 (in Chinese with English abstract).

[15] Singhal A, Seborg DE. Pattern matching in multivariate time series databases using a moving window approach. Ind. Eng. Chem. Res., 2002,41(16):3822−3838.

[16] Krzanowski WJ. Between-Groups comparison of principal components. Journal of the American Statistical Association, 1979, 74(367):703−707.

[17] Guan HS, Jiang QS, Wang SJ.A new similarity measure for clustering multivariate time series. Journal of Computational Information Systems, 2007,3(5):2031−2036.

Address for correspondence:

Name: Yuanjian Zhang
Yu Li
Address: 235,East Nanjing Road,Nanchang,Jiangxi, 330047, P.R.China
Email 708912734@qq.com
ly4232@hotmail.com