Health Sensors Information Processing and Analytics Using Big Data Approaches

D. Gachet Páez(✉), M.L. Morales Botello, E. Puertas, and M. de Buenaga

Universidad Europea de Madrid, 28670 Villaviciosa de Odón, Spain
{gachet,mariadelaluz.morales,enrique.puertas,buenaga}@uem.es

Abstract. In order of maintain the sustainability of the public health systems it is necessary to develop new medical applications to reduce the affluence of chronic and dependent people to care centers and enabling the management of chronic diseases outside institutions Recent advances in wireless sensors technology applied to e-health allow the development of "personal medicine" concept, whose main objective is to identify specific therapies that make safe and effective individualized treatment of patients based for example in remote monitoring. The volume of health information to manage, including data from medical and biological sensors make necessary to use Big Data and IoT concepts for an adequate treatment of this kind of information. In this paper we present a general approach for sensor's information processing and analytics based on Big Data concepts.

Keywords: Big data · Internet of things · Cloud computing · Elderly · Sensors

1 Introduction

The rapidly growing popularity of health care and activity monitoring applications for smart mobile devices like smart phones and tablets provide new ways to collect information about people health status, both manually and automatically. Also, there are appearing new COTS (*Commercial Off-The-Shelf)* wearable medical sensors that can easily connect with the smart phones or tablets via Bluetooth and transfer the sensing measures directly to a public or private cloud infrastructure. This has provided a more efficient and convenient way to collect personal health information like blood pressure, oxygen saturation, blood glucose level, pulse, electrocardiogram (ECG), etc., that can be analyzed for generating alarms or furthermore, it would also be possible to track the patient's behaviors on a real-time basis and over long periods, providing a potential alert for signs of physical and/or cognitive deterioration [1].

Medical and bio-signal sensors are also commonly used in Intensive Care Units (ICU) at hospitals and the information provided by them can be used for example to develop methods for patient-specific prediction of in-hospital mortality. Sensors used in ICUs can provide precise, heterogeneous and continuous information about clinical condition of a patient as for example heart rate, invasive mean arterial blood pressure, invasive diastolic arterial blood pressure, invasive systolic arterial blood pressure, etc. All this data can be processed and analyzed in order to predict special clinical situation or as above mentioned in-hospital mortality.

B. Mandler et al. (Eds.) IoT 360° 2015, Part I, LNICST 169, pp. 481–486, 2016.
DOI: 10.1007/978-3-319-47063-4_52

2 Big Data Processing and Analysis

One of the most important aspects when we are dealing with health monitoring is how the data generated by the sensor and medical devices is processed and analyzed. The first thing we have to think about is the goal, that is, we need to establish what we want to do before thinking about how we are going to achieve it. Health data mining approaches are similar to standard data mining procedures, and is performed basically in five stages [2].

Data acquisition and preprocessing. The three most important data sources are experimental data, public datasets, and simulated data. In the first scenario, data is usually gathered from a set of wearable devices that are monitoring a group of test users. Public datasets are those that have been made publicly available in sites like UCI ML Repository or Kaggle.com. When data is gathered from many heterogeneous wearable devices or sources, a normalization of data step is required. Data preprocessing involves data cleaning for removing noise and data interpolation for mitigate the effects of missing values.

Data Transformation. When there are a big number of attributes, dimensionality reduction is a required step because it improves efficiency and reduces over fitting. There are usually two ways to do this task: feature selection and feature extraction [3].

Modeling. This stage, applies knowledge discovery algorithms to identify patterns in the data or predict some variables, at this point we can apply several algorithms as for example Rule induction learners, Decision trees, Probabilistic Learners, Support Vector Machines (SVM), Hidden Markov Models (HMM), etc.

Evaluation. The effectiveness of learning algorithms systems is measured in terms of the number of correct and wrong decisions. Some of the metrics used for evaluating the modes are recall and precision. Recall is defined as the proportion of class members assigned to a category by a classifier. Precision is defined as the proportion of correctly assigned documents to a category.

3 Processing Cardiovascular Data

As use case for data processing and analytics we use real data set obtained from Physionet Computing in Cardiology Challenge 2012: Predicting Mortality of ICU Patients [4]. The origin of data were the hospital medical information systems for Intensive Care Unit (ICU) patients with ICU stays lasting at least 48 h. The dataset consisted of Set A and Outcome-related descriptors (csv) text file. Set A was composed of four thousand records (text files) corresponding to the four thousand ICU stays (patients), and each record was composed of up to 37 time series variables (such as Heart Rate, Weight, pH, SysABP, DiasABP, Urine, …) which could be observed once, more than once, or not at all in some cases (not at all records), and could be recorded at regular intervals (hourly, daily) or at irregular intervals. The time stamps of the measurement indicated the elapsed time since admission to the ICU. In addition to the previous variables, each record included six general descriptors collected at the time the patient was admitted in ICU (RecordID, Age, Gender, Height, ICUType, and Weight). These descriptors appeared at the beginning of each record (time 00:00). In correspondence with Set A, the Outcomes was a file composed by four thousand rows, where each row contained six outcome-related descriptors for each record (patient). These descriptors were: RecordID, SAPS-I score [5], SOFA score [6], Length of stay (days), Survival (days) and In-hospital death (0 indicated survival and 1 indicated in-hospital death).

All valid values for general descriptors, time series variables, and outcome-related descriptors were non-negative *(≥0).* A value of −1 indicated missing or unknown data. The four thousand records of individual patients that make up the Set A were joined together resulting in a file next to 2 million of rows (1885594 rows). This amount of data cannot be processed by many conventional analysis tools. In order to process and analyze this big amount of data, we used R, an open source software for statistical computing [7]. In this work, we used the tools of R to perform a predictive model from the cardiovascular data described previously.

3.1 Predictive Modelling with R

The aim of the model is to predict in-hospital mortality (0: survival, or 1: in-hospital death) of each patient from the corresponding variables and descriptors. The first step for building a predictive model about the patient's mortality in UCU is to perform data formatting and pre-processing. The text file of 1885594 rows (and four columns: RecordID, Time, Variable, Value), which contained the complete time series variables of the four thousand patients, was saved as a "data table" in R. This allowed us to process big data with high speed. In order to get static variables, that is, in order to work with a unique value for each time series variable, for each patient (record) we calculated the median of the measurements for each variable.

Then, we constructed a structure where the (37) variables are the columns and the 4000 patients (RecordID) are the rows. As well as, we added 8 columns corresponding to the following general and outcomes-related descriptors: RecordID, Age, Gender, Height, ICUType, SAPS.I, SOFA and InHospitalDeath.

Before using these data for the logistic regression model, we carried out a data pre-processing, which consisted of:

1. Replacing invalid physiological values with valid values in the descriptor Height (for example, height value of 13 cm probably corresponds to 130 cm).
2. Assigning NA (Not Acknowledge) to both outlier and invalid values of the following variables: pH, NISysABP, NI DiasABP, DiasABP, MAP (for example, a value of 0 in NISysABP);
3. Replace −1 with NA from missing or unknown data (which were indicated with −1 in the original dataset).

Logistic regression is a common analysis technique for situations with binary outcome data [8, 9]. This method has been employed by several participants of the Physionet Challenge 2012 to produce predictions of the binary variable "InHospital-Death" [10, 11]. In this work, we used the same method to predict survival or in-hospital death using the statistical software R. The logistic regression was performed using the function "glm" included in "stats" package of R. The dependent variable of the model was InHospitalDeath and the independent variables will be the rest of the columns (preprocessed variables and descriptors) previously presented. Due to the logistic model in R deletes the missing observations, we only used as independent variables the variables or descriptors in which missing data were present in less than 10 % of patients. In addition, we deleted the rows (patients) with missing data in any column (796 of 4000 rows). We applied repeatedly the model in order to select the more significant variables, and only the variables with a statistical significance level ($p < 0.001$) were included in the final model. For training the logistic regression model we used the 60 % of the patients (training dataset, 1922 patients) and the remaining 40 % (testing dataset, 1282) was used to test the model. The variables finally considered for inclusion in the logistic regression model are presented in Table 1.

Table 1. The first column shows the variables that were selected for the model according to a few missing values. The second column shows the variables finally used by the model based on the higher significance.

All Variables	Model Var.
BUN Blood urea nitrogen	X
Creatinine	
GCS Glasgow Coma Score	X
Glucosa Serum glucose	X
HCO3 Serum bicarbonate	
HCT Hematocrit	X
HR Heart rate	X
K Serum potassium	
Lactate	
Mg Serum magnesium	X
Na Serum sodium	
Platelets	
Temp Temperature	X
Urine	
Urine.Sum	X
WBC White blood cell count	
Weight	
Age	X
Gender	
ICUType	
SAPS.I	
SOFA	

We used the function “predict” included in “stats” package in R to predict the probability of death of the testing dataset patients using the model obtained with the training dataset patients. The predicted outcome is a value between 0 and 1. In order to get a binary outcome, that is, to predict survival (0) or in-hospital death (1), we assigned 0 to the predicted value when the probability predicted was lower than 0.5, and in otherwise, we assigned 1 to the predicted value.

For model evaluation we take into account the official metric used for Physionet Challenge 2012, score 1 (s1), defined as the minimum value between Sensitivity (Se) and Positive Predictivity (P^{+}):

$$S_e = \frac{TP}{TP + FN} \tag{1}$$

$$P^{+} = \frac{TP}{TP + FP} \tag{2}$$

TP is the number of true positives, FP is the number of false positives and FN is the number of false negatives. True positive indicates that the model predicts 1 when InHospitalDeath is 1, false positive indicates that the model predicts 1 when InHospitalDeath is 0, and false negative indicates that the model predict 0 when InHospitalDeath is 1. Therefore, the Se value quantifies the fraction of in-hospital deaths that are predicted, and P^{+} quantifies the fraction of correct predictions of in-hospital deaths.

4 Conclusion and Future Work

We obtained a fraction of correct predictions of in-hospital deaths, P^{+} of 0.455. The fraction of in-hospital deaths that are predicted, Se, was 0.006. Therefore, the score s1 obtained by our model was of 0.006. Our fraction of correct predictions was higher than the s1 value obtained by the winners of Challenge 2012 [12] using Set A. This result suggests us that the fraction of correct predictions of in-hospital deaths given by our model is relatively good. However, the fraction of in-hospital deaths predicted by our model was small. A possible cause is that many variables (22 physiological variables) were not taken into account by the model due to frequent missing data. However, frequent missing data does not imply a minor relation between the variables and the patient death. An information gain analysis performed between the median of each variable and the in-hospital death variable (results not shown) revealed that variables which were in the variable group with longer weights, i.e., the variables better related with the death of the patient (such as PaCo2, Bilirubin, Albumin and AST) were rejected by high missing observations.

Other aspects that could affect to the results of our model are the diverse population with a wide variety of life-threatening conditions, with frequent missing and occasionally incorrectly recorded observations, idiosyncrasies of care administration, and highly unbalanced class sizes that make up the dataset. Whatever the cause, our logistic regression model can be improvable, however, the aim of this work was not to get the best model, but carry out a R implementation of a predictive model based on cardiovascular (big) data. Despite of poor performance, the methodology proposed in this research using the statistical package R can be used for analyzing other biomedical datasets. R has thousands of libraries that can help to analyze and visualize complex datasets, and it lets researchers to deal with big data, providing libraries and functions for cleaning and analyzing large volumes of data produced by medical devices and sensors.

Acknowledgments. This work is still being developed with funds granted by the Spanish Ministry of Economy and Competitiveness under project iPHealth (TIN-2013-47153-C3-1).

References

1. Fundación Vodafone: Innovación TIC para las personas mayores. Situación, requerimientos soluciones en la atención integral de la cronicidad y la dependencia (2011). http://www.vodafone.es/static/fichero/pro_ucm_mgmt_015568.pdf
2. Sow, D.M., Turaga, D.S.: Schmidt: mining of sensor data in healthcare: a survey. In: Aggarwal, C.C. (ed.) Managing and Mining Sensor Data, pp. 459–504. Springer, Berlin (2013)
3. Apiletti, D., Baralis, E., Bruno, G., Cerquitelli, T.: Real-time analysis of physiological data to support medical applications. Trans. Inf. Tech. Biomed. **13**, 313–321 (2009)
4. Physionet 2012 Cardiovascular Challenge. http://physionet.org/challenge/2012/
5. Le Gall, J.R., Loirat, P., Alperovitch, A., Glaser, P., Granthil, C., Mathieu, D., Mercier, P., Thomas, R., Villers, D.: A simplified acute physiology score for ICU patients. Crit. Care Med. **12**(11), 975–977 (1984)
6. Ferreira, F.L., Bota, D.P., Bross, A., Mélot, C., Vincent, J.L.: Serial evaluation of the patients. JAMA **286**(14), 1754–1758 (2001)
7. R project: http://www.r-project.org/
8. Hosmer, D.W., Lemeshow, S.: Applied Logistic Regression, 2nd edn. Wiley, New York (2000)
9. Hamilton, S.L., Hamilton, J.R.: Predicting in-hospital-death and mortality percentage using logistic regression. Comput. Cardiol. **39**, 489–492 (2012)
10. Vairavan, S., Eshelman, L., Haider, S., Flowers, A., Seiver, A.: Prediction of mortality in an intensive care unit using logistic regression and hidden Markov model. Comput. Cardiol. **39**, 393–396 (2012)
11. Bera, D., Nayak, M.M.: Mortality risk assessment for ICU patients using logistic regression. Comput. Cardiol. **39**, 493–496 (2012)
12. Johnson, A.E.W., Dunkley, N., Mayaud, L., Tsanas, A., Kramer, A.A., Clifford, G.D.: Patient specific predictions in the intensive care unit using a Bayesian ensemble. Comput. Cardiol. **39**, 249–252 (2012)
13. GachetPáez, D., Aparicio, F., de Buenaga, M., Ascanio, J.R.: Big data and IoT for chronic patients monitoring. In: Hervás, R., Lee, S., Nugent, C., Bravo, J. (eds.) UCAmI 2014. LNCS, vol. 8867, pp. 416–423. Springer International Publishing, Cham (2014)
14. Sahoo, S.S., Jayapandian, C., Garg, G., Kaffashi, F., Chung, S., Bozorgi, A., et al.: Heart beats in the cloud: distributed analysis of electrophysiological big data using cloud computing for epilepsy clinical research. J. Am. Med. Inform. Assoc. **21**(2), 263–271 (2014)
15. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. **41**, 15:1–15:58 (2009)