ASIAN JOURNAL OF PHARMACEUTICAL AND CLINICAL RESEARCH

INNOVARE ACADEMIC SCIENCES
Knowledge to Innovation

Special Issue (April)

Online - 2455-3891
Print - 0974-2441

Full Proceeding Paper

PREDICTIVE ANALYTICS OF HEALTH-CARE DATA

APURVA WAGHMARE, SWEETLIN HEMALATHA*

School of Computer Science and Engineering, VIT University, Chennai, Tamil Nadu, India. Email: sweetlin.hemalatha@vit.ac.in

Received: 19 January 2017, Revised and Accepted: 20 February 2017

ABSTRACT

Predictive analytics is employed to improve the ability to take precautionary measures during medical emergencies. In health care, the sensor-based data are generated daily which can be used to predict future data using regression model. In this paper, pain dataset from integrating data for analysis, anonimyzation, and sharing repository is used for experimenting different machine algorithms. The results show that logistic regression gives more accuracy than other algorithms.

Keywords: Predictive analytics, Machine learning, Naïve Bayes, Logistic regression, Random forest.

 DOI: http://dx.doi.org/10.22159/ajpcr.2017.v10s1.19750

INTRODUCTION

In health care, sensors are worn on the body to measure vital parameters of the body. In this way, the monitoring of the patient can be improved and the received data may be stored for future use. These sensors send the sensed phenomenon at regular time intervals. The data thus collected are used for analyzing the patterns and predict the future values. If the reading of the sensors goes out of order or any anomaly occurs, then alert message can be sent. Since mobiles are omnipresent, they may be used to alert the patient and doctor [1]. In turns, the doctor can give proper medication and this will prove very helpful in critical cases. In this paper, algorithms such as linear regression model, Naive Bayes (NB) algorithm, and random forest (RF) are used to predict the pain experienced by the patients in intensive care unit (ICU) as "pain" or "no pain."

TECHNIQUES USED

Logistic regression (LR)

In this algorithm, there is one dependent variable suppose "Y" and there are "n" number of independent variables "$X_1, X_2, X_3 ...., X_n$." As pain is being predicted, then Y is the pain parameter, and the independent variables are the sensor data measured in the ICU. These parameters are explained in this paper in section III. Then, a model can be created by calculating the coefficients of independent variables, i.e., $a_1, a_2, a_3, a_4 ....a_n$. The LR equation is formed as follows:

$$Y=a_1X_1+a_2X_2+a_3X_3+.......+a_nX_n$$

This equation is the regression model which can be used for prediction of the dependent variable Y [2].

In LR, the dependent variable can take only two values. Since pain is being predicted, therefore Y can be concluded as pain or no pain. If Y value is equal to one, then it can be interpreted as the patient is suffering through pain. If Y value is equal to zero, then it can be interpreted as the patient is not suffering through pain.

To find the probability P of predicted value, the following formula is used:

$$P=\exp(Y)/(\exp(Y)+1)$$

The resultant P value is the probability of Y.

NB algorithm

In this algorithm, the best suitable conditions for class label to be true are found using posterior probability method. The algorithm requires the class label value to be categorical. For each parameter, find values such that the probability of class value to be true is the highest. In posterior probability, this calculation is independent of the other parameters which might be correlated with each other. This classifier algorithm thus finds the best possible value to favor the class label to be true and creates a model using which the value of class label can be predicted. Whenever data are to be predicted, then each parameter is checked whether the given data fit the prediction model or not [3].

As prediction of pain is the aim of the model, the class label is categorical, i.e., either there is pain or no pain. Thus, NB algorithm is tried to form a model for prediction.

RF algorithm

In RF algorithm, the data are divided into number of subsets depending on rows. Rows are selected randomly to form subsets and for each subset a decision tree is formed. Now suppose a dataset is divided into n number of subsets and then N number of decision trees are generated. Depending on the data each subset contains, the decision tree model will be varying. Now, if any datum is fed to this algorithm, then all the subsets will give their conclusions of class label value depending on the model built. The class label voted maximum by the decision trees, is the solution [4].

The data used for prediction of pain are sensor data. When RF is applied to this data set, then any row will randomly go to any subset. The sequence of sensor data according to time will be lost.

DATA SET

To do analysis, the selected dataset is of pain from iDASH repository that stands for "integrating data for analysis, anonimyzation, and sharing." The data set does not reveal the identity of patient and also it is asynchronous with respect to time of taking observations. The data are generated by sensors and there are many vital parameters such as heart rate, blood pressure, and laboratory test results. The dataset contains data generated for 53 patients followed at USCD Medical Centre ICUs [5].

The dataset for each patient contains 24 parameters. They are as explained below:

1. Central parenteral nutrition (CPN) Glasgow coma scale score: It has three tests such as eye, verbal, and motor scale.
2. Skin Braden scale activity: It is a tool for predicting patients' risk for developing pressure ulcers.
3. Point-of-care testing (POCT) glucose test result: POCT glucose test

is a medical diagnostic testing at bedside of patient.

4. CPN Glasgow coma scale best motor response: This is measured after four stages, i.e., check, observe, stimulate, and rate.
5. Motor response in left lower extremity, left upper extremity, right lower extremity, and right upper extremity: All are measured to check the control and reflex movements and thus abnormality can be detected.
6. CPN Glasgow coma scale best verbal response: Verbal response can be oriented, confused, words, sounds, none.
7. CPN Glasgow coma scale eye opening: Eye-opening response is spontaneous, to sound, to pressure, none.
8. Blood pressure: Measurement of pressure between heartbeats.
9. Skin Braden scale score: For predicting pressure ulcer risk.
10. Skin Braden scale friction and shear: Measure of the amount by how much the patient needs to be moved. Shear means the skin and bone can move in opposite direction to cause breakdown.
11. Skin Braden scale sensory perceptions: To detect and respond to discomfort or pain that is related to pressure on that part.
12. Ventilator tidal volume exhaled: It is set at respiratory rate and used when patient is not able to trigger ventilator.
13. John Hopkins fall risk total: Total of many variables such as age, fall history, and mobility to measure fall risk.
14. Nursing FiO_2: It is the oxygen delivery to the patient.
15. Pupil size left: Unequal pupil size may be or may not be a symptom of underlying disorder.
16. Mean arterial pressure (MAP) cuff: MAP is the measure of average blood pressure in one cycle.
17. Skin Braden scale moisture and nutrition: Skin is exposed to moisture to which degree and usual food intake pattern, respectively.
18. Pulse: The tactile arterial palpation of the heartbeat by trained fingertips.
19. Richmond Agitation-Sedation Scale score: Measures level of consciousness.
20. Respiration: Number of breaths per minute.
21. Ventilation respiratory rate: The rate at which breaths occur is called Ventilation respiratory rate. It is measured in breaths per minute. This parameter of the patient was also observed.
22. Pulse oximetry: Used to measure level of oxygen in the blood without inserting any instrument in the body.
23. Temperature: Body temperature.
24. Pain: Measured on a scale of 0-9, where 0 stands for no pain.

PROPOSED SYSTEM

Most of the existing health-care analytics focus on classification of health data streams and deciding the situation is normal or abnormal. But that detection may be too late to control the situation of risk. Hence, this research focuses mainly on developing a prediction of stream mining system that analyses the track of the health data streams and provides a probabilistic solution.

The patients in ICU are monitored continuously using various sensors that generate sensor data [6]. The proposed model is such that depending on the data collected for each patient, a personalized LR model should be generated which will be able to predict pain for that patient. More the data, better will be the prediction. This will be helpful for predicting the pain experienced by the patient and doctors can be informed to take required measures.

Sensor on patient's body

There are sensors worn on patient's body to measure vital parameters. They are used for continuous monitoring. The sensors sense the parameters and send the data continuously or after fixed time intervals.

Fig. 1 depicts the flow of data for predicting the class variable "pain" as "pain" or "no pain."

Patient's sensor data

The sensors send data continuously. This datum is stored and is also used for predicting pain [7]. There are supposing N number of patients in ICU sending data. This datum is used to build a LR model to predict pain.

Prediction model for patients

This model is generated using LR. The model predicts whether the patient will have pain or not. The result of this model is further sent to take actions accordingly.

Is pain predicted?

This block monitors the outputs of all the patients' prediction models and checks if any patient is predicted to be suffering through pain. If any patient is going to have pain, then alert message will be sent to the doctor to take precautionary measure. If pain is not predicted, then the analysis is continued.

RESULTS AND DISCUSSION

IBM SPSS is the software which is used for statistical analysis. This software is used to generate regression model for the given dataset. It generates the coefficients of each independent variable. Furthermore, it shows the percentage of accuracy achieved in prediction model and Pearson correlation model also.

Weka is the software and it is a collection of machine learning algorithms used for data mining tasks.

Figs. 2 and 3 show the performance of three techniques, namely, LR, NB, and RF in terms of accuracy and time, respectively. The confusion

Table 1: Comparison of confusion matrix

A	B	Classified as
LR		
31	94	Pain
17	870	No pain
NB		
52	73	Pain
114	773	No pain
RF		
31	94	Pain
17	870	No pain

LR: Logistic regression, NB: Naive Bayes, RF: Random forest

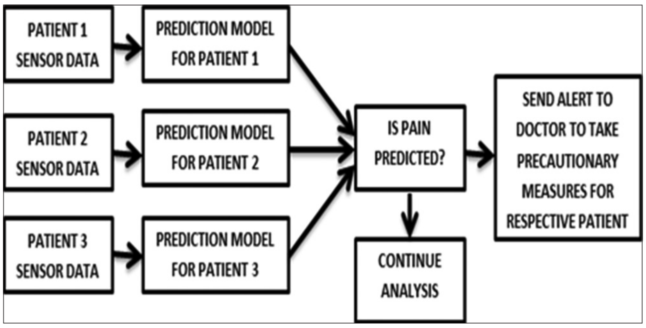


Fig. 1: Block diagram of pain prediction system

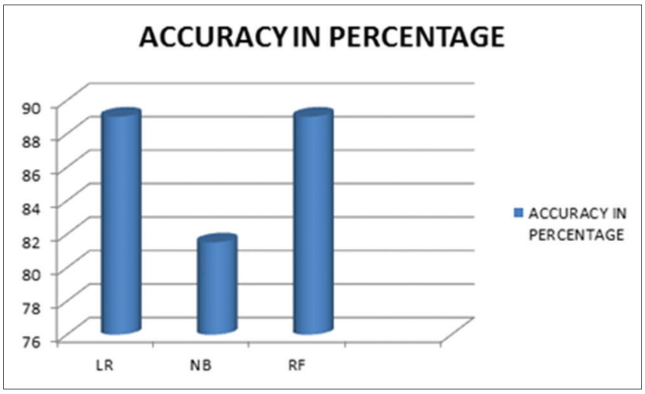


Fig. 2: Comparison result of accuracy of different techniques used

matrices of three techniques are presented in Table 1.

LR model is the best among all the three algorithms explained earlier.

LR is better than NB algorithm for prediction because in NB the independent variables are not correlated with each other. The health parameters are always correlated. The increase or decrease in one parameter may cause variations in other parameters also. Hence, we conclude that LR model is the better than NB algorithm for prediction of pain.

LR is better than RF because RF takes more time. In health care, time plays an important role and a little delay in decision may result in increase in critical condition of patient or may result fatal.

Verifying the performances of these algorithms in Weka for dataset of one patient following results is obtained.

LR gives a better confusion matrix as compared to NB. RF is confusion matrix similar to LR in this case. However, this is not true for every patient. Although with respect to accuracy RF is close to LR but with respect to time, it is not.

Thus, LR model is the best among all the three algorithms.

LR model for the first patient,

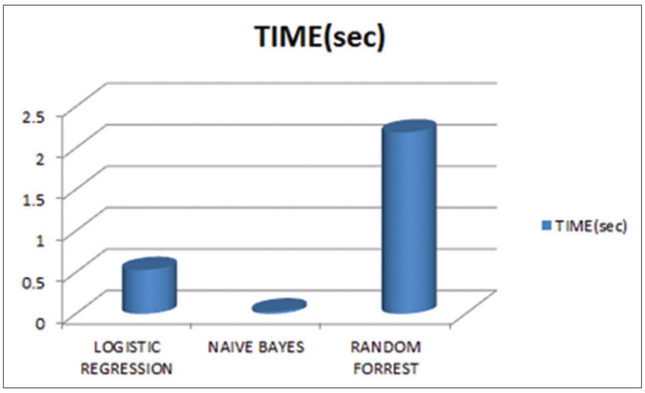
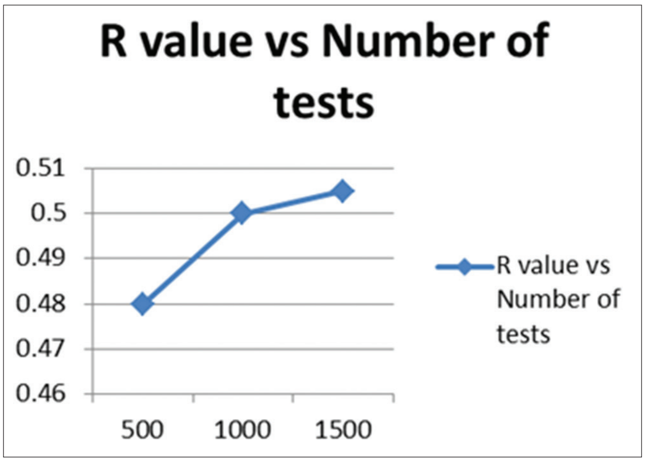


Fig. 3: Comparison result of time of execution of different techniques used



Patient 1

R value versus number of tests
0.48 versus 500
0.5 versus 1000
0.505 versus 1500

Fig. 4: R value versus number of tests for patient 1

$$Y=-0.128X_1-0.298X_2-0.063X_3-0.006X_4-0.070X_5+0.363X_6-0.377X_7-0.366X_8+0.553X_9+0.212X_{10}-0.027X_{11}+0.367X_{12}-0.194X_{13}+0.133X_{14}-0.140X_{15}+0.050X_{16}-0.140X_{17}-0.042X_{18}+0.038X_{19}-0.099X_{20}0.030X_{21}+\ 0.191X_{22}-0.004X_{23}+0.323X_{24}-0.006X_{25}+0.002X_{26}-0.093X_{27}-0.048X_{28}-0.073X_{29}$$

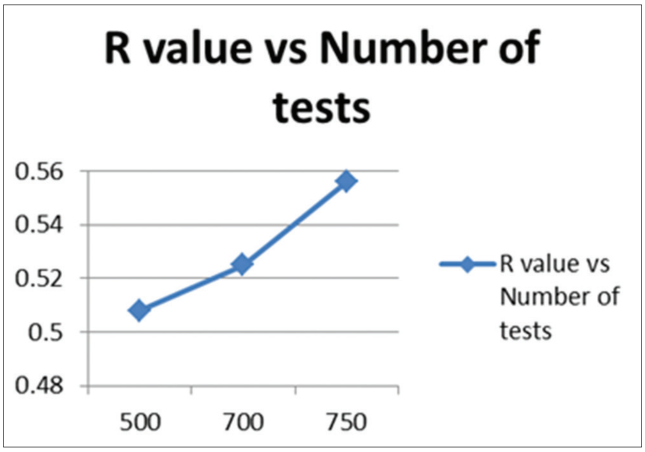
$$P=\exp(Y)/(\exp(Y)+1)$$

Figs. 4-7 show the results obtained by analyzing the sample dataset in IBM SPSS software.

It is evident from the experimental results that the model has to be tuned for an individual as the R value varies from patient to patient even for the same number of tests sampled for different patients. Moreover, it is observed that with increasing number of records, the accuracy of prediction increases.

CONCLUSION AND FUTURE WORK

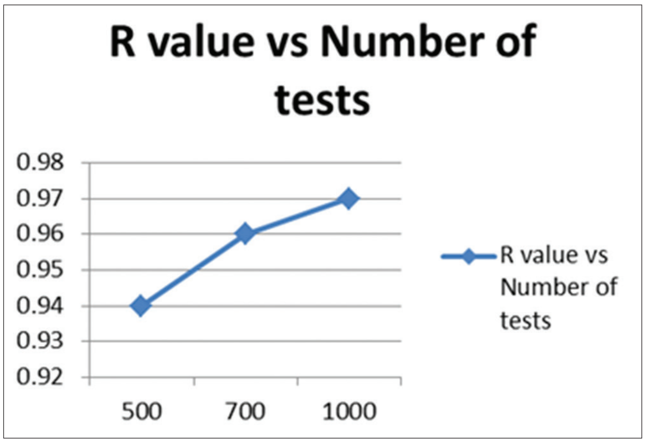
This paper presents the results of analysis of prediction models for foretelling the presence/absence of pain felt by patients in ICU.



Patient 2

R value versus number of tests
0.508 versus 500
0.525 versus 700
0.556 versus 750

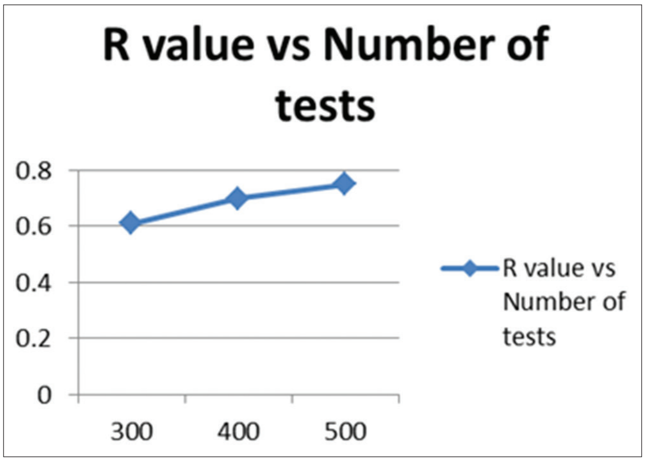
Fig. 5: R value versus number of tests for patient 2



Patient 3

R value versus number of tests
0.94 versus 500
0.96 versus 700
0.97 versus 1000

Fig. 6: R value versus number of tests for patient 3



Patient 4

R value versus number of tests
0.61 versus 300
0.7 versus 400
0.75 versus 500

Fig. 7: R value versus number of tests for patient 4

Experimental results show that LR gives better accuracy and in less time compared to other techniques for the dataset considered. Although there is a need for personalized prediction model, the current work focuses on prediction model built on static data set. Hence, the future work concentrates on building an incremental prediction model for the streaming data.

REFERENCES

1. Mukherjee A, Pal A, Misra P. Data analytics in ubiquitous sensor-based health information systems. In: Next Generation Mobile Applications, Services and Technologies (NGMAST). 6th International Conference on IEEE, September; 2012. p. 193-8.
2. Howbert J. Machine Learning Logistic Regression; 2012. Available from: https://www.search.datacite.org/data-centers/cdl.idashrep?query=pain#
3. Naive Bayes Classifier; 2009. Available from: https://www.en.wikipedia.org/wiki/Naive_Bayes_classifier.
4. Ho TK The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 1998;20(8):832-44.
5. Wang S. Pain Prediction Data; 2014. Available from: https://www.search.datacite.org/data-centers/cdl.idashrep?query=pain#.
6. Poh N, Tirunagari S, Windridge D. Challenges in designing an online healthcare platform for personalised patient analytics. In: Computational Intelligence in Big Data (CIBD), IEEE Symposium on IEEE, December; 2014. p. 1-6.
7. Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I. Spark: Cluster computing with working sets. In: Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing. Vol. 10. June; 2010. p. 10.