APAMI2020 General Oral Presentation Session | APAMI 2020 | General Oral Presentation Session
Artificial Intelligence in Healthcare and Clinical Practice (2)
Mon. Nov 23, 2020 9:00 AM - 10:10 AM Room E-1 (Congress center 5F - Conference Room 52)

[AP3-E1-1-03] An Analysis on Remote Healthcare Data for Future Health Risk Prediction to Reduce Health Management Cost

*Shaira Tabassum[1], Masuda Begum Sampa[2], Rafiqul Islam Maruf[3], Fumihiko Yokota[4], Naoki Nakashima[3], Ashir Ahmed[2] (1. Department of Computer Science and Engineering, United International University, Bangladesh, 2. Department of Advanced Information Technology, Kyushu University, Japan, 3. Medical Information Center, Kyushu University Hospital, Japan, 4. Institute of Decision Science for a Sustainable Society, Kyushu University, Japan)

Keywords: Remote Healthcare, Portable Health Clinic (PHC), Data Preprocessing, Categorical Variables, Encoding

Machine Learning (ML) is tremendously enhancing the healthcare sector by continuously collaborating in diverse healthcare circumstances. It explores thousands of data beyond human capability, analyzes medical conditions, and suggests outcomes with clinical insights. Still, a large segment of the population does not get quality healthcare due to insufficient medical facilities and socio-economic conditions. Thus, the Portable Health Clinic (PHC) aims to make technology-enabled smart healthcare affordable to the unreached population. The paper reports a comparative analysis of seven supervised learning algorithms to predict the possible health risk in future using the remote healthcare data provided by PHC. The survey and clinical data of PHC have been used in this work to predict the triage condition that a patient may have later. Four categorical variable encoding and two missing value handling techniques have been applied for preprocessing and preparing the healthcare data. The preprocessed PHC data has achieved 90.34% accuracy on the Random Forest Classifier. Thus, this pre-informing health service will reduce health management costs and allow people to take necessary mitigating actions to minimize health risks.

©Japan Association for Medical Informatics

An Analysis on Remote Healthcare Data for Future Health Risk Prediction to Reduce Health Management Cost

Shaira Tabassum[a], Masuda Begum Sampa[b], Rafiqul Islam Maruf[c], Fumihiko Yokota[d], Naoki Nakashima[c] and **Ashir Ahmed[b]**

[a] *Department of Computer Science and Engineering, United International University, Bangladesh*
[b] *Department of Advanced Information Technology, Kyushu University, Japan*
[c] *Medical Information Center, Kyushu University Hospital, Japan*
[d] *Institute of Decision Science for a Sustainable Society, Kyushu University, Japan*

Abstract

Machine Learning (ML) is tremendously enhancing the healthcare sector by continuously collaborating in diverse healthcare circumstances. It explores thousands of data beyond human capability, analyzes medical conditions, and suggests outcomes with clinical insights. Still, a large segment of the population does not get quality healthcare due to insufficient medical facilities and socio-economic conditions. Thus, the Portable Health Clinic (PHC) aims to make technology-enabled smart healthcare affordable to the unreached population. The paper reports a comparative analysis of seven supervised learning algorithms to predict the possible health risk in future using the remote healthcare data provided by PHC. The survey and clinical data of PHC have been used in this work to predict the triage condition that a patient may have later. Four categorical variable encoding and two missing value handling techniques have been applied for preprocessing and preparing the healthcare data. The preprocessed PHC data has achieved 90.34% accuracy on the Random Forest Classifier. Thus, this pre-informing health service will reduce health management costs and allow people to take necessary mitigating actions to minimize health risks.

Keywords:

Remote Healthcare, Portable Health Clinic (PHC), Data Preprocessing, Categorical Variables, Encoding, Missing Data, Machine Learning, Supervised Learning Algorithm

Introduction

Artificial Intelligence (AI) and Machine Learning (ML) are lending hands in healthcare and revolutionising the medical industry around the world. The multitudes of increasing ML applications aim to provide value-based care to millions of people. ML has the potential to process millions of datasets beyond the scope of human capability, analyse the medical data, and provide better outcomes with clinical insights to aid physicians [1]. Even after this tremendous revolution, more than a billion people do not get access to quality healthcare service. A large segment of this population are low-income people from rural areas of developing countries [2].

Long-distance patient and physician contact, telehealth or telemedicine, is an effective way to overcome this barrier to provide quality care to the people of rural and remote areas [3]. The PHC is a compact telehealth system which aims to provide a smart healthcare system to the low-income and unreached people of remote areas. This system has a PHC device box containing different medical sensors, a group of certified health workers, an online database to preserve health data, and a panel of doctors to remotely serve patients. The online database contains past health records of patients for monitoring them remotely and future health analysis purposes. It has introduced a triage system to classify the overall health condition of the patient based on the collected health data. The data is collected from the same subjects in every three months. Each round of data has individual triage values. This triage value determines patients' physical condition based on the survey data and clinical data of the PHC health dataset.

In this work, we have predicted the second-phase triage status of patients using their first-phase health data with machine learning models. So, the system is taking the present health status information of a patient and analyzing it to predict upcoming possible health risks. The dataset contains 46 distinct features. We have preprocessed the dataset with four encoding methods (one-hot encoding, label encoding, frequency encoding, and one-hot-frequency encoding) to convert the categorical variables to numerical values. The missing data has been handled using two techniques (listwise deletion, and mean imputation). Based on the results of these preprocessing methods on seven algorithms, a comparative analysis is presented in the results section. Our prediction system has achieved 90.34% accuracy on Random Forest Classifier with frequency encoding.

Most of the people around the world are diagnosed after getting noticeable signs of illness. As a result, it becomes too late and also very costly to treat the disease. Therefore, early health risk prediction enables patients and doctors to take precautionary steps soon before symptoms appear. Patients can receive early treatment before it becomes a severe disease. Thus, we expect this work will create an affordable and sustainable health service for the middle and low-income people of unreached areas. People will be pre-informed about their health conditions and take emergency physician consultation which will reduce health management costs and minimize possible health risks.

Preliminaries

This section presents a brief background of different categorical variable encoding techniques and supervised learning algorithms related to this research.

Handling Missing Data

- Listwise Deletion: Delete all instances with any missing value. This method is supported for a large number of procedures were missing values appear at random positions [4].

- Mean Imputation: This method is known as one of the measurements of center tendency. It replaces all the missing values with the mean value of that variable [4].

Categorical Variable Encoding Techniques

Most of the statistical machine learning models do not accept categorical values as input. Thus, for using categorical variables in machine learning purposes, several techniques are used to encode them into numerical variables [5].

- One-hot Encoding: One-hot encoding takes a single variable with n instances and d distinct values and transforms it to d binary variables, each containing n instances. Each instance is represented with binary values where 1 indicates presence and 0 indicates absence of the variable [5]. One-hot encoding makes sure that the machine learning model does not give priority to the higher numbers.
- Label Encoding: Label encoding is simply assigning categorical values to integer values. For d distinct values, it converts the categorical values in [0, d-1] range. This technique shows poor performance to some machine learning models as it creates artificial numerical orders. So the model considers that the higher numbers are more important [6].
- Frequency Encoding: This encoding technique utilizes the frequency of categories as labels. First, it groups by the categorical variable to count the number of appearances of each category. Then, divides it with the total number of instances to calculate frequency of each category. Thus, the category with highest frequency is prioritized and its somewhat related with the target variable [7].
- One-hot-frequency Encoding: This technique is a combination of one-hot encoding and frequency encoding. It applies frequency encoding if the categorical variable has 1-5 distinct values and one-hot encoding if the categorical variable has more than five distinct values [8].

Supervised Learning Algorithms

- K-Nearest Neighbours (KNN): In KNN algorithm, K is the number of neighbours closest to the query that are considered for voting. KNN can generate different classification outputs from the same inputs with different values of K [9].
- Support Vector Machine (SVM): SVM maps n number of features into an n-dimensional feature space. Then it identifies hyperplanes to separate the features maintaining maximum marginal distance and minimum classification errors [10].
- Artificial Neural Network (ANN): ANN appears to be an interconnected group of nodes. The output of one node is the input of another node [9].
- Logistic Regression (LR): LR finds probability to determine whether an instance belongs to a certain class or not. It assigns an instance to a class if its probability value is higher than a threshold value [9].
- Linear Discriminant Analysis (LDA): It projects the higher dimension space features onto the lower dimension space. It calculates between-class variance and within-class variance to construct lower dimensional space called Fisher's criterion [11].
- Decision Tree (DT): DT models create a tree-like structure considering the decision logics based on the input variables. The internal nodes represent different tests and the leaf or terminal node determines the decision outcome [12].
- Random Forest (RF): RF consists of many DTs like a forest with a variety of collection of trees. The DTs of a RF are trained on different parts of the training dataset. Thus, each DT provides different classification outcomes. So the forest chooses the final outcome with the average (numeric outcome) or most votes (discrete outcome) of all trees. As RF considers outcomes of many DTs, it can reduce the variance from a single DT on the same dataset [9].

Portable Health Clinic System and Healthcare Related Data Collection

A PHC is an eHealth system that aims to provide affordable primary healthcare services to prevent severity or to control non-communicable diseases (NCDs).

Portable Health Clinic System Architecture

PHC system has been jointly developed by Grameen Communications, Bangladesh, and Kyushu University, Japan as a tele-medicine system for the unreached communities with a special focus on non-communicable diseases [2]. This system consists of four components, as in Figure 1:

1. PHC Device Box: It contains various medical sensors, internet enabled tablet pc and a printer.
2. Health Worker: A certified healthcare worker with micro entrepreneurship training.
3. GramHealth Database System: An Online data-server for sharing and preservation of health data.
4. Doctors Call Center: A panel of doctors to consult with remote patients and to provide e-Prescriptions.

The clinical measurements taken by a PHC are as follows:

(1) blood pressure (2) pulse rate (3) body temperature (4) oxygenation of blood (SpO2) (5) arrhythmia (6) body mass index (BMI) (7) waist, hip, and W/H ratio, (8) blood glucose (9) blood cholesterol (10) blood hemoglobin (11) blood uric acid (12) blood grouping (13) urinary sugar, and (14) urinary protein.

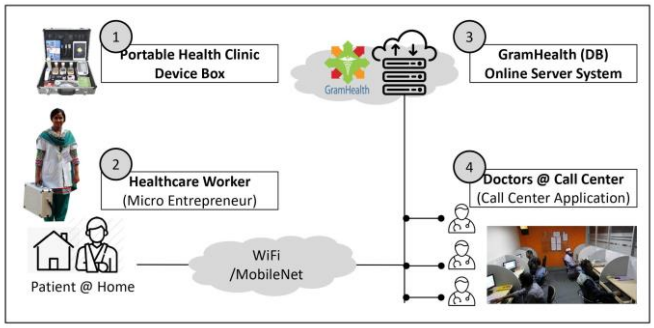

Figure 1- Four components of PHC system

A health worker visits a patient with the PHC box to measure the vital information and upload this data with medical history of the patient to an online server using the GramHealth Client application. The remote doctor gets access to this data and makes a video call to the patient for further verification. Finally, the doctor creates an online prescription and preserves it to the online server under each patient's profile. The health worker accesses the system, prints the prescription from the server and passes it to the patient with detailed explanation instantly. The whole process takes about 15 to 30 minutes per patient.

The PHC system introduces a triage system to classify the subjects in four categories, namely, (i) Green or Healthy (ii) Yellow or Suspicious (iii) Orange or Affected and (iv) Red or Emergent, based on the gradual higher risk status of health. The subjects under Orange and Red who are primarily diagnosed as in the high risk zone need a doctors' consultation.

Data Collection

Five different healthcare data were collected from 271 corporate employees working in Grameen Bank Complex. A cross-sectional survey was conducted in August 2018 among all office workers who agreed to participate in PHC health checkups and eHealth services in the Grameen Bank Complex in Dhaka, Bangladesh. The Grameen Bank Complex holds several different offices, such as Grameen Bank, Grameen Communications, other non-government organizations, and private companies, with more than 500 workers. This study recruited participants from these 18 institutions.

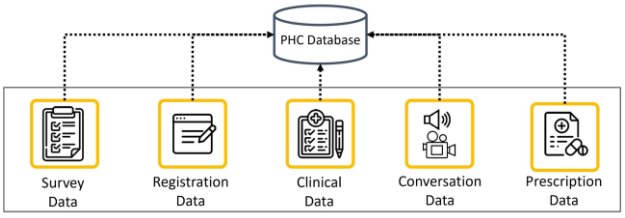

Figure 2- Inside PHC-DB: Data were collected in five different steps

The characteristics of the data are described below.

1. Survey Data: Diet, Nutrition, Physical activities, Mental state etc. Structured.
2. Registration Data: Name, Age, Location, Mobile No, User ID, Checkup ID, Site ID etc. Structured.
3. Clinical Data: Height, Weight, BMI, Waist, Hip·PBS/FBS·SPO2, Temp, BP etc. Structured.
4. Conversation Data: Audio data. Doctor-patient conversation. Multi-language. Unstructured.
5. Prescription Data: Medication, Advice, Dr info, Health status. Unstructured.

Methods

The research has been conducted by preprocessing the collected PHC dataset with several techniques to handle categorical variables and missing values and finally, analysis and comparison among seven machine learning models to predict future health status of patients.

Overview of future health prediction system

Initially the collected PHC Dataset has been preprocessed with four categorical variable encoding methods (one-hot, label, frequency, and one-hot-frequency) and two missing data handling methods (listwise deletion, and mean imputation). The preprocessed dataset is then fed to seven different statistical machine learning models such as KNN, NDA, DT, RF to predict the future health status of the patient. An overview of the whole process is shown in Figure 3.

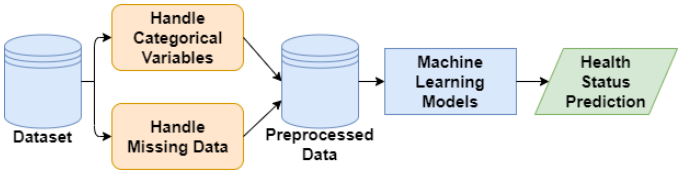

Figure 3- Block diagram of future health prediction system

The five-steps PHC health data is collected from the same subjects every three months. Each round of data has individual triage value which represents the current health status of the subject. Currently, PHC has 271 patients' health data in four-phases. In this work, the survey data and clinical data of the first two phases of PHC health data have been used to predict the future health risk of a patient. This work has predicted the triage value of the second phase using the health data of the first phase. Thus, based on the current physical conditions, we can predict the future health risk of a patient.

Data Preprocessing

There are 271 subjects' health data in the first phase. Among them, 115 subjects' data was collected in the second phase. Therefore, listwise deletion and mean imputation have been applied for handling the missing data.

Table 1- Encoding outcomes on PHC health data

Encoding Approaches	Encoded Categorical Attributes	Total Attributes
One-hot Encoding	495	511
Label Encoding	30	46
Frequency Encoding	30	46
One-hot-frequency Encoding	78	93

Table 2- Future health risk prediction outcomes on PHC health data

Machine Learning Algorithms	One-hot Encoding		Label Encoding		Frequency Encoding		One-hot-frequency Encoding	
	Listwise Deletion	Mean Imputation	Listwise Deletion	Mean Imputation	Listwise Deletion	Mean Imputation	Listwise Deletion	Mean Imputation
Random Forest	75.12%	90.09%	74.76%	89.4%	75.73%	**90.34%**	73.88%	90.18%
Decision Tree	31.43%	69.14%	48.57%	65.43%	38.24%	**72.84%**	40%	**72.84%**
Logistic Regression	40%	**77%**	37%	75%	41%	**77%**	34%	**77%**
K-Nearest Neighbour	40%	**76.54%**	45.71%	71.6%	50%	71.6%	45.71%	74.07%
Support Vector Machine	42.86%	77.78%	42.86%	77.78%	47.06%	79.01%	42.86%	**80.25%**
Artificial Neural Network	68%	76%	56%	74%	54%	79%	62%	**80%**
Linear Discriminant Analysis	34.29	**71.6%**	34.29%	65.43%	38.24%	66.67%	37.14%	**71.6%**

Future Health Risk Prediction Using Machine Learning Models

After preprocessing the data, seven popular supervised learning algorithms (GNB, KNN, SVM, ANN, LR, LDA, DT, RF) have been applied to predict the future triage status of the patient. The results and comparisons are demonstrated in the next section.

Results and Analysis

To predict the future health risk in the best possible way, we have applied several encoding and missing data handling techniques on seven machine learning models. The comparison is given in Table 2.

In our analysis, RF with frequency encoding and mean imputation has outperformed all other approaches with 90.34% accuracy. One-hot and one-hot-frequency encoding have also achieved very similar accuracy with RF classifier. Mean imputing has performed well, whereas listwise deletion has shown very poor performance in all cases. This happened due to the small size of the training set. As listwise deletion discards all the instances with missing values, it has trained the model with a very small dataset. As a result, the model is overfitted on the training set and does not generalize well on the test set [13].

Among the encoding techniques, one-hot-frequency encoding has achieved top accuracy in five algorithms. In RF and KNN, its accuracy is very similar to the highest score achieved technique. In our paper [8], one-hot-frequency outperformed other encoding techniques in analyzing medical data. However, one-hot and frequency encoding also performed well on several algorithms. On the other hand, label encoding could not beat other techniques in any of the approaches.

Comparing the algorithms performances, RF has achieved highest accuracy in all cases including listwise deletion. Ali et al. has shown that RF classifiers perform well both on large and small datasets [14]. Some other works also demonstrated RF as the top algorithm to analyze medical data [8], [15].

Conclusion

Portable Health Clinic (PHC) is a technology-enabled telehealth system which aims to provide affordable health service to the people of rural and remote areas. The online PHC database regularly stores patients' health data after a certain period of time in different phases and classifies the medical condition of a patient as a triage system in four categories. This system determines whether a patient falls under a high risk zone and needs emergency doctor consultation or not. We have predicted the future health risk possibilities (second-phase triage status) based on the current physical condition (first-phase health data) of a patient. To make a more efficient prediction, we have preprocessed the data through several encoding and missing value handling techniques. Applying this technique on RF classifier, frequency encoding has achieved 90.34% accuracy which is the highest score. One-hot-frequency has got top accuracies on five algorithms. It has achieved 90.18% accuracy with RF which is very similar to the highest accuracy score. Hence, this pre-informed health service approach enables people to take required treatment in proper time, which will cure different health hazards and decrease health management costs.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 18K11529.

References

[1] Corbett E. The real-world benefits of machine learning in healthcare. Health Catalyst. 2017.

[2] Ahmed A, Inoue S, Kai E, Nakashima N, Nohara Y. Portable Health Clinic: A pervasive way to serve the unreached community for preventive healthcare. InInternational Conference on Distributed, Ambient, and Pervasive Interactions. 2013 Jul 21; pp. 265-74. Springer, Berlin, Heidelberg.

[3] EE GB, Farr R, Ben Raimer MD. Benefits of telemedicine in remote communities & use of mobile and wireless platforms in healthcare.

[4] Cheema JR. A review of missing data handling methods in education research. *Review of Educational Research*. 2014 Dec; 84(4):487-508.

[5] Potdar K, Pardawala TS, Pai CD. A comparative study of categorical variable encoding techniques for neural network classifiers. *International journal of computer applications*. 2017 Oct; 175(4):7-9.

[6] Hancock JT, Khoshgoftaar TM. Survey on categorical data for neural networks. *Journal of Big Data*. 2020 Dec; 7:1-41.

[7] Roy B. All about categorical variable encoding. Towards DataScience. 2019.

[8] Tabassum S, Sampa MB, Islam R, Nakashima N, Ahmed A. A data enhancement approach to improve machine learning

performance for predicting health status using remote healthcare data (submitted).

[9] Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*. 2019 Dec; 19(1):1-6.

[10] Joachims T. Making large-scale SVM learning practical. Technical Report. 1998.

[11] Sawla S. Linear discriminant analysis. Medium Data Science. 2018.

[12] Quinlan JR. Induction of decision trees. *Machine learning*. 1986 Mar 1; 1(1):81-106.

[13] Ying X. An overview of Overfitting and its solutions. InJournal of Physics: Conference Series. 2019 Feb 1 (Vol. 1168, No. 2, p. 022022). IOP Publishing.

[14] Ali J, Khan R, Ahmad N, Maqsood I. Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*. 2012 Sep 1; 9(5):272.

[15] Naseer S, Saleem Y. Enhanced Network Intrusion Detection using Deep Convolutional Neural Networks. *TIIS*. 2018 Oct 1; 12(10):5159-78.

Address for correspondence

Shaira Tabassum
Department of Computer Science and Engineering
United International University, Bangladesh
E-mail: stabassum152129@bscse.uiu.ac.bd