S & M 1646

Implementation of Integrated Electronic Health Record and Mobile Personal Health Record Datasets for Improving Healthcare Services

Sol-Bee Lee,[1] Jung-Hyok Kwon,[1] Eui-Jik Kim,[1*] and Jaehoon Park[2**]

[1]Department of Convergence Software, Hallym University,
1 Hallymdaehak-gil, Chuncheon-si, Gangwon-do 24252, South Korea
[2]Department of Electronic Engineering, Hallym University,
1 Hallymdaehak-gil, Chuncheon-si, Gangwon-do 24252, South Korea

(Received March 30, 2017; accepted April 12, 2018)

Keywords: decision tree, EHR, healthcare service, integrated dataset, mPHR

Medical big data are rapidly being generated and accumulated throughout the healthcare industry. Using such medical data to extract meaningful information is expected to improve healthcare services significantly. In this paper, we present an integrated dataset, consisting of electronic health records (EHRs) and mobile personal health records (mPHRs), which enables high-accuracy disease diagnostics. An EHR represents the overall health status of a patient, including the patient's past medical records, while mPHRs are data recorded by an individual's mobile devices and provide real-time health information that varies over time. Accordingly, each EHR and mPHR plays a complementary role in diagnosing a patient's disease, enabling accurate health diagnostic services. To generate an integrated dataset comprising EHR and mPHR, two tasks are performed for each individual dataset: data preprocessing and data matching. The former includes a formatting step to set the appropriate format for the data and a cleansing step to replace missing values and outliers with median values. The latter task requires overwriting or combining individual attributes within the EHRs and mPHRs into a unified form. For a comparative analysis of the integrated datasets, we generate a prediction model for heart disease using the decision tree method. The results show that the prediction model for the integrated dataset exhibits higher accuracy than individual datasets in predicting a patient's disease.

1. Introduction

Recently, a large amount of medical data previously stored in the healthcare industry in hard copy formats has been rapidly digitized and accumulated.(1) In addition, with the increasing prevalence of mobile devices along with the rapid development of medical devices and platforms based on the Internet of Things (IoT), medical big data collection for individual users has become possible.(2,3) Many organizations and hospitals can obtain valuable information from these digitized, vast collection of electronic health records (EHRs). In recent years,

[*]Corresponding author: e-mail: ejkim32@hallym.ac.kr
[**]Corresponding author: e-mail: jaypark@hallym.ac.kr

https://dx.doi.org/10.18494/SAM.2018.1896

ISSN 0914-4935 © MYU K.K.
http://myukk.org/

various medical studies using EHRs have been conducted to obtain meaningful information.(2) In other words, awareness of the importance of medical big data analytics is spreading. Thus, technologies such as data mining should be used to find meaningful information in vast amounts of data effectively.

The EHRs used in most medical institutions to improve the quality of care and the health of a patient consist of health information related to the patient's overall health care and are recorded electronically.(4) Medical big data such as EHRs are difficult to analyze by conventional methods owing to the variety of data types and the existence of high-dimensional data. Furthermore, EHRs are relatively static compared with mobile personal health records (mPHRs); thus, it is difficult to determine a patient's current health status using EHRs. Nonetheless, EHRs contain the care and prescription information received from health professionals and have the advantage that they enable health providers to provide patients with a higher quality of care.

On the other hand, mPHRs are collected through mobile healthcare devices. A variety of sensors attached to mobile healthcare devices detect the data needed for healthcare such as blood pressure, body temperature, and electrocardiogram, which constitute the mPHR. It is difficult to verify data the accuracy of mPHRs because their data are not collected by medical experts or using medical devices at a hospital. Therefore, mPHRs are difficult to use for in-depth analysis related to a disease. However, since the patient's status is measured and updated over a relatively short period compared with EHRs, there is an advantage in that the current status of the patient is more closely reflected. To address the limitations of the two datasets and provide improved healthcare services, extensive studies related to EHRs and mPHRs have been conducted in various aspects.

Yan *et al.* proposed a novel hybrid outlier detection method called pruning-based K-nearest neighbor (PB-KNN).(5) This method integrates density-based, cluster-based methods and KNN algorithms to overcome the difficulties of analyzing data in medical fields with a large amount of data, a variety of data types, and high-dimensional data. It aims to effectively reduce data dimensions and detect outliers in large-scale healthcare data that cannot be easily processed using conventional methods. However, the proposed method is limited for analysis since it uses only EHRs. Hauskrecht *et al.* proposed a new data-driven monitoring and alerting framework.(6) The purpose of the framework is to detect clinical outliers using past patient cases stored in EHRs. To create an outlier detection model, it uses the time series data in the EHRs that represent a patient's status observed at each interval of time. However, it is difficult to identify the actual status of a patient using the framework because it is based on previously recorded EHRs. Bouri *et al.* identified the benefits and challenges of mPHRs and provided recommendations for their support and use.(7) mPHRs are portable and interoperable digital files that allow patients to manage and transmit information. The patient's status can be identified through their mPHRs in the event of an emergency because the mPHRs can be accessed via the Internet. However, the authors discussed only the necessity and application of mPHRs. Therefore, experiments and analyses related to mPHRs need to be conducted.

In this paper, we present an integrated dataset for healthcare services consisting of EHRs and mPHRs. The integrated dataset was generated using EHRs from the heart disease dataset of the UCI machine learning repository combined with experimentally generated mPHRs.(8) Using the EHRs, we can identify a patient's past medical history, vital signs, medications, radiology reports, and other data.(9) The mPHR is effective in helping patients manage disease by allowing them to track their health status outside a medical institution.(7) The integrated dataset can provide an accurate health diagnosis for a patient and enables access to complementary information from the EHR and mPHR when diagnosing a patient's disease. In addition, it focuses on the existence of heart disease. To generate the integrated dataset, we conducted two tasks: data preprocessing and data matching. Data preprocessing involves data formatting and cleansing. We set the appropriate data formats during formatting and replaced missing values and outliers with median values during cleansing. Then, we performed data matching to overwrite or combine individual attributes within the EHRs and mPHRs into a unified form for the dataset. To demonstrate the superiority of the integrated dataset, a decision tree was used to perform a comparative analysis of the EHRs, PHRs, and the integrated dataset to assess their accuracies when used in diagnosing heart disease. The results show that the decision tree for the integrated dataset is 3 and 4% more accurate than the EHRs and mPHRs, respectively, in diagnosing heart disease.

The rest of this paper is organized as follows. In Sect. 2, we describe the integrated EHR and mPHR dataset. The experiment and performance analysis for the integrated dataset using R version 3.4.1 are presented in Sect. 3. Finally, we conclude this paper in Sect. 4.

2. Integrated EHR and mPHR Dataset

In this section, we describe the creation of an integrated dataset using EHRs and mPHRs. The purpose of the integrated dataset is to diagnose the existence of heart disease more accurately by reflecting the current status of the patient. We created the EHRs using data from the heart disease dataset directory in the UCI machine learning repository to perform an analysis focused on heart disease. The heart disease dataset directory contains four datasets related to the diagnosis of heart disease. Each dataset is collected from one of four locations (i.e., Cleveland Clinic Foundation, OH, USA; Hungarian Institute of Cardiology, Budapest, Hungary; Veterans Affairs Medical Center, CA, USA; and University Hospital, Zurich, Switzerland) and consists of multiple attributes, some of which may not be related to heart disease. Thus, to identify only the attributes associated with heart disease, we sorted the attributes according to their frequency of appearance in the dataset and extracted the attributes with a higher frequency than the predefined value. In this paper, we set the predefined value to 0.5, and 14 attributes with the same instance format were extracted. Table 1 shows the 14 attributes of the EHRs. The values of all attributes of the EHRs are expressed as numerical values. We integrate these four datasets collected from different locations and use them as the EHRs.

The EHRs consist of 920 records that facilitate a higher quality of care. These records include many attributes associated with heart disease, which are measured at medical institutions. However, EHRs are relatively static compared with mPHRs because the data in the former are updated only when the patient receives medical care at a hospital. If the analysis is performed using only EHRs, it would be difficult to diagnose a patient's current health status precisely. Therefore, we integrate the EHRs with mPHRs to overcome this limitation. mPHRs

Table 1
Description of EHR attributes.

Attribute	Description	Example of value
age	age in years	63
sex	sex (1 = male, 0 = female)	1
cp	chest pain type (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)	1
trestbps	resting blood pressure	145
chol	serum cholesterol	233
fbs	fasting blood sugar > 120 mg/dl (1 = true, 0 = false)	1
restecg	resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy)	2
thalach	maximum heart rate achieved	150
exang	exercise-induced angina (1 = yes, 0 = no)	0
oldpeak	ST depression induced by exercise relative to rest	2.3
slope	the slope of the peak exercise ST segment (1 = positive slope, 2 = flat, 3 = negative slope)	3
ca	number of major vessels colored by fluoroscopy	0
thal	thalassemia (3 = normal, 6 = fixed defect, 7 = diagnosis of heart disease)	6
num	existence of heart disease (0 = no, 1 = yes)	0

are datasets measured periodically by a mobile healthcare device. Using mPHRs for analysis can reveal the current status of a patient. In our work, mPHRs are experimentally generated to match the attributes of the EHRs. The mPHRs consist of four cardiovascular-disease-related attributes that can be collected via a mobile healthcare device. These four attributes are blood pressure, heart rate, blood sugar, and activity calories. The mPHRs are also composed of 920 records, similar to the EHRs.

To create the integrated dataset, we perform data preprocessing and data matching for each individual record in the EHR and mPHR. Data preprocessing is conducted for each record in the two datasets and is an essential task because missing values and outliers in a dataset reduce the accuracy of analysis. More accurate analysis can be performed by eliminating missing values and outliers from a dataset through data preprocessing. We replace missing values and outliers in the EHRs with the median value of the corresponding attribute. Then, an integrated dataset containing the EHRs and mPHRs is generated by data matching. Data matching is the process of replacing the values of attributes in EHRs with those of attributes in mPHRs and adding the attributes to the EHRs if the attributes contained only in mPHRs exist. This data matching process is repeated for all patients listed in the EHRs since each mPHR maintains the information for a single patient. As a result, if the attributes of the mPHRs are the same as those of the EHRs, the values of attributes are overwritten in the EHRs. Otherwise, the attributes from the mPHRs are combined with those of the EHRs. The integrated dataset consists of 15 attributes and 920 records related to heart disease expressed as numerical values. The integrated dataset is periodically updated from the mPHR to determine a patient’s current status.

3. Experiments

In this section, we describe and discuss the experimental results in detail. The experiment was conducted using R version 3.4.1. For comparative analysis of the integrated datasets, we generated a prediction model for heart disease using a decision tree, which provides IF-THEN rules for data classification. In the experiment, we classified the existence of heart disease in the patients into ‘Yes’ or ‘No’ using the ‘num’ attribute.

We compared the accuracy of the prediction model based on the decision tree generated from the integrated dataset containing the EHRs and PHRs. Figure 1 shows the decision tree for the integrated dataset. In the figure, the decision tree has a total of 15 rules. With these rules, the existence of heart disease is classified as ‘Yes’ or ‘No’. The existence of heart disease for all patients in the integrated dataset is predicted through the decision tree. For example, in the first rule associated with ‘cp’, if the value of ‘cp’ for a certain patient is not 1, 2, or 3, the health data moves to the second rule associated with ‘trestbps’. In the second rule, if the value of ‘trestbps’ is smaller than 132, the health data moves to the third rule. In the third rule associated with ‘chol’, if the data is greater than 430, the existence of heart disease is predicted as ‘Yes’. Figure 2 shows the decision tree for the EHRs. This decision tree is similar to that of the integrated dataset in that ‘cp’ is the first rule and the data is segmented. However, it has fewer rules compared with the decision tree for the integrated dataset, and thus the accuracy of the prediction model is lower. The decision tree for the mPHRs is shown in Fig. 3. In the figure, the decision tree generated from the mPHRs has only five rules because the mPHRs contain only four attributes related to cardiovascular disease.

Table 2 shows the accuracy of each decision tree. Each decision tree was analyzed to assess the accuracy of the predicted results derived from the decision tree. We set *n* and *p* for the accuracy calculation as follows: *n* is the number of data points with a prediction value of ‘No’ when the actual value is ‘No’, and *p* is the number of data points predicted as ‘Yes’ when the actual value is ‘Yes’. The accuracy of the decision tree can be given as (*n* + *p*)/(the number of total records of the dataset). The results show that the decision tree for the integrated dataset has the highest accuracy (i.e., 0.82) among the three datasets. Specifically, the accuracy is 3 and 4% higher than that of the EHRs and mPHRs, which have accurate predictive values of 0.79 and 0.78, respectively.

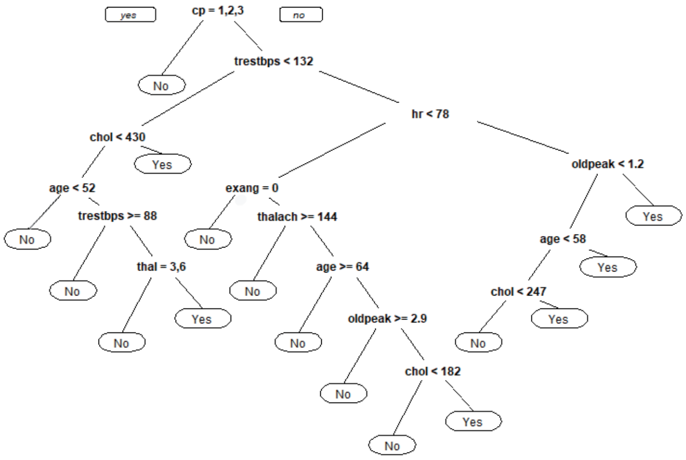

Fig. 1. Decision tree for integrated dataset.

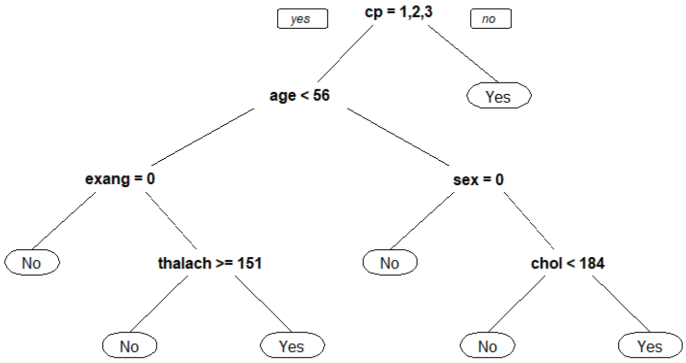

Fig. 2. Decision tree for EHRs.

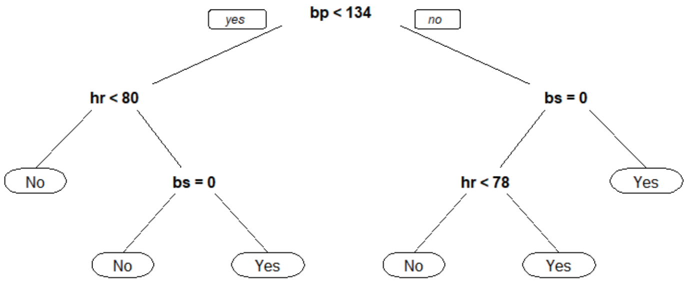

Fig. 3. Decision tree for mPHRs.

Table 2
Accuracy of decision trees.

	Integrated dataset	EHRs	mPHRs
Accuracy	0.82	0.79	0.78

4. Conclusions

In this paper, we present an integrated dataset for healthcare services consisting of EHRs and mPHRs. To create the integrated dataset, data preprocessing and data matching were conducted. Data preprocessing involves formatting and cleansing, during which the appropriate format for the data was specified and the missing values and outliers were replaced with median values. Data matching was used to overwrite or combine individual attributes within the EHRs and mPHRs into a unified form for the dataset. An experiment was conducted for comparative analysis of the integrated datasets. For this, we generated a prediction model using the decision tree to predict the existence of heart disease in patients. The results show that the integrated datasets are 3 and 4% more accurate than the EHRs and mPHRs, respectively, in predicting the existence of heart disease. We expect that our work will serve as a reference for creating an integrated dataset for more effective and accurate healthcare in the near future.

Acknowledgments

This research was supported by Hallym University Research Fund, 2018 (H20180248).

References

1 W. Raghupathi and V. Raghupathi: Health Inform. Sci. Syst. **2** (2014) 1.
2 D. V. Dimitrov: Healthcare Inf. Res. **22** (2016) 156.
3 P. K. Sahoo, S. K. Mohapatra, and S.-L. Wu: IEEE Access **4** (2016) 9786.
4 Office of the National Coordinator for Health Information Technology (ONC): https://www.healthit.gov/buzz-blog/electronic-health-and-medical-records/emr-vs-ehr-difference/ (accessed July 2017).
5 K. Yan, X. You, X. Ji, G. Yin, and F. Yang: Proc. 2016 IEEE Int. Conf. Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom) (IEEE, 2016) 157.
6 M. Hauskrecht, I. Batal, M. Valko, S. Visweswaran, G. F. Cooper, and G. Clermont: J. Biomed. Inform. **46** (2013) 47.
7 N. Bouri and S. Ravi: JMIR Mhealth Uhealth **2** (2014) 1.
8 University of California: http://archive.ics.uci.edu/ml/datasets/heart+Disease (accessed July 2017).
9 N. Menachemi and T. H. Collum: Risk Manage. Healthcare Policy **4** (2011) 47.