Journal of Infection and Public Health (2016) **xxx**, xxx—xxx

ELSEVIER

http://www.elsevier.com/locate/jiph

Building predictive models for MERS-CoV infections using data mining techniques

Isra Al-Turaiki, Mona Alshahrani*, Tahani Almutairi

Information Technology Department, College of Computer and Information Sciences,
King Saud University, Saudi Arabia

Received 23 June 2016; received in revised form 20 July 2016; accepted 6 September 2016

KEYWORDS
MERS-CoV;
Data mining;
Decision tree;
J48;
Naive Bayes;
Classification

Summary

Background: Recently, the outbreak of MERS-CoV infections caused worldwide attention to Saudi Arabia. The novel virus belongs to the coronaviruses family, which is responsible for causing mild to moderate colds. The control and command center of Saudi Ministry of Health issues a daily report on MERS-CoV infection cases. The infection with MERS-CoV can lead to fatal complications, however little information is known about this novel virus. In this paper, we apply two data mining techniques in order to better understand the stability and the possibility of recovery from MERS-CoV infections.

Method: The Naive Bayes classifier and J48 decision tree algorithm were used to build our models. The dataset used consists of 1082 records of cases reported between 2013 and 2015. In order to build our prediction models, we split the dataset into two groups. The first group combined recovery and death records. A new attribute was created to indicate the record type, such that the dataset can be used to predict the recovery from MERS-CoV. The second group contained the new case records to be used to predict the stability of the infection based on the current status attribute.

Results: The resulting recovery models indicate that healthcare workers are more likely to survive. This could be due to the vaccinations that healthcare workers are required to get on regular basis. As for the stability models using J48, two attributes were found to be important for predicting stability: symptomatic and age. Old patients are at high risk of developing MERS-CoV complications. Finally, the performance of all the models was evaluated using three measures: accuracy, precision, and recall. In general, the accuracy of the models is between 53.6% and 71.58%.

* Corresponding author.
E-mail addresses: ialturaiki@ksu.edu.sa (I. Al-Turaiki), monaalshahrani@outlook.com (M. Alshahrani), 435203979@student.ksu.edu.sa (T. Almutairi).

http://dx.doi.org/10.1016/j.jiph.2016.09.007

Conclusion: We believe that the performance of the prediction models can be enhanced with the use of more patient data. As future work, we plan to directly contact hospitals in Riyadh in order to collect more information related to patients with MERS-CoV infections.

Introduction

In 2012, Saudi Arabia witnessed the outbreak of a virus called Middle East Respiratory Syndrome Coronavirus (MERS-CoV). The novel virus belongs to the coronaviruses family which is responsible for causing mild to moderate colds. MERS-Co is blamed for causing severe acute respiratory illness that lead to death in many cases. According to [1], MERS-CoV symptoms include: cough, fever, nose congestion, breath shortness, and sometimes diarrhea. The virus began spreading rapidly in Saudi Arabia in 2013. Since then, the Control and Command Center of Saudi Ministry of Health in Saudi Arabia started recording and reporting the cases. The ministry website provides daily statistics on new confirmed MERS-CoV cases, recoveries, and deaths.

Infection with MERS-CoV can lead to fatal complications. Unfortunately, there is little information about how the virus spreads and how patients are affected. Data mining is the exploration of large datasets to extract hidden and previously unknown patterns and relationships [2]. In healthcare, data mining techniques have been widely applied in different applications including: modeling health outcomes and predicting patient outcomes, evaluation of treatment effectiveness, hospital ranking, and infection control [3].

In this paper, we build several models to predict the stability of the case and the possibility of recovery from MERS-CoV infection. The goal is to better understand which factors contribute to complications of this infection. The models are built by applying data mining techniques to the data provided by the Control and Command Center of Saudi Ministry of health website [1].

The rest of the paper is organized as follows. In *Literature review* section, we review related work in the applications of data mining in healthcare. *Methodology* section describes the dataset, pre-processing steps, data mining techniques, and our experimental results. Finally, *Conclusion* section concludes the paper with findings.

Literature review

In this section, we highlight some of the related work in data mining applications in healthcare.

Data mining has been widely used for the prognosis and diagnoses of many diseases. Ferreira et al. [4] used data mining to improve the diagnosis of neonatal jaundice in newborns. In their experiment, the dataset consisted of 70 variable collected for 227 healthy newborns. Many data mining techniques were applied, including: J48, CART, Naive Bayes classifier, multilayer perceptron, SMO, and simple logistic. The best predictive models were obtained by using Naive Bayes, multilayer perceptron, and simple logistic. For heart disease diagnoses, Venkatalakshmi and Shivsankar [5] compared the performance of decision tree algorithm and Naive Bayes. The experimental results using a dataset of 294 records with 13 attributes showed that the performance of the two algorithms is comparable. FP-growth, Association rule mining, and decision trees were used for the diagnosis and prognosis of breast cancer [6]. The classification models were built using a dataset of 699 records and 9 attributes and the best accuracy was achieved using decision trees induction algorithms.

In terms of survivability predicting, Bellaachia et al. [7] used Naive Bayes, back-propagated neural network, and the C4.5 decision tree algorithm to predict the survivability of breast cancer patients. The dataset used in the study was obtained from the Surveillance Epidemiology and End Results (SEER). Experimental results indicated that the C4.5 algorithm outperformed the other two techniques. Recently, several predictive models for breast cancer survival were developed [8]. The models were based on a dataset of 657,712 records and 72 variables, also obtained from SEER. Three different

Please cite this article in press as: Al-Turaiki I, et al. Building predictive models for MERS-CoV infections using data mining techniques. J Infect Public Health (2016), http://dx.doi.org/10.1016/j.jiph.2016.09.007