Journal of Infection and Public Health (2016) **9**, 744—748

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].

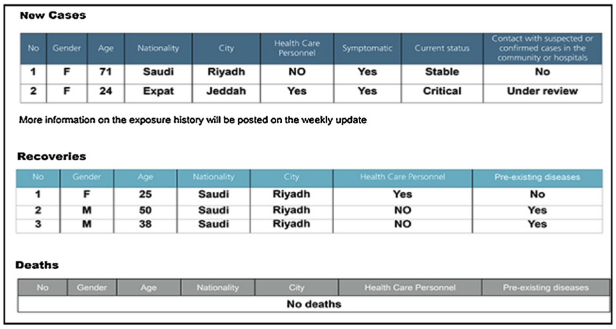
The rest of the paper is organized as follows. In *Literature review* section, we review related work in the applications of data mining in healthcare. *Methodology* section describes the dataset, pre-processing steps, data mining techniques, and our experimental results. Finally, *Conclusion* section 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 data mining techniques were used: Support Vector Machine (SVM), Bayes Networks, and Chi-squared Automatic Interaction Detection (CHAID). Results showed that the best survival prediction model was obtained using SVM. The authors in [8] presented a study of predictive models for breast cancer survival. The main goal was to discover important attributes that contribute to breast cancer survival. Three data mining techniques were used: Support Vector Machine (SVM), Bayes Net, and Chi-squared Automatic Interaction Detection (CHAID). Experiments on a dataset obtained from SEER showed that the SVM model outperformed other models in terms of accuracy, sensitivity, and specificity. SVM was able to identify ten attributes that are important indicators of breast cancer survivability. Sandhu et al. [9] proposed a cloud-based MERS-CoV prediction system. The system is based on Bayesian Belief Networks (BBN) for initial classification of patients. A geographic positioning system is utilized to represent patients on Google Maps. Patients classified as infected were tracked using GPS from their mobile phones. The proposed system is useful to citizens since it allows them to avoid infected areas. In addition, healthcare authorities can manage the infection problem more effectively. The BBN achieved an accuracy of 83.1% on synthetic data.



New Cases

No	Gender	Age	Nationality	City	Health Care Personnel	Symptomatic	Current status	Contact with suspected or confirmed cases in the community or hospitals
1	F	71	Saudi	Riyadh	NO	Yes	Stable	No
2	F	24	Expat	Jeddah	Yes	Yes	Critical	Under review

More information on the exposure history will be posted on the weekly update

Recoveries

No	Gender	Age	Nationality	City	Health Care Personnel	Pre-existing diseases
1	F	25	Saudi	Riyadh	Yes	No
2	M	50	Saudi	Riyadh	NO	Yes
3	M	38	Saudi	Riyadh	NO	Yes

Deaths

No	Gender	Age	Nationality	City	Health Care Personnel	Pre-existing diseases
No deaths						

Figure 1 Sample of the original dataset.

Methodology

Dataset description and pre-processing

As mentioned earlier, our dataset was obtained from the website of the Control and Command Center of Saudi Ministry of Health [1]. We used the data on MERS-CoV infections reported between 2013 and 2015. The data was published in three separate categories: new cases, recoveries, and deaths. For all the categories, the following patient information was provided: gender, age, nationality, city, and whether the patient is a healthcare personnel or not. In addition, there was more specific information for each category. The additional information is as follows:

- New cases: symptomatic, current status, and whether the patient had any contact with suspected or confirmed MERS-CoV infection case.
- Recoveries and deaths: does the patient have pre-existing diseases.

We collected 633 new case records, 231 recovery records, and 218 death records, for a total of 1082 records. A sample of the original dataset is shown in Fig. 1. The dataset was published in different formats. Some of the records were found as text files. Other records were provided in image format. Thus, our first step was to prepare the data in a unified format appropriate for data mining. Information in image format was manually extracted. Records with missing and inconsistent values (ex. gender is adult) were excluded. The age attribute was converted to discrete values. Finally, all records were prepared in .csv format. In order to build our prediction models, we split the dataset into two groups. The first group consisted of 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.

Data mining

Classification is a widely used technique in healthcare. Here, we build several classification models to predict the stability and recovery of MERS-CoV infection. We apply Naive Bayes and J48 [10] decision tree algorithm. Here, we briefly describe these algorithms.

- Naive Bayes classifier: is a probabilistic model based on Bayes theorem. It assumes class conditional independence, where the dependencies between class attributes are ignored. Research has shown that Naive Bayes classifiers have comparable performance to other classification algorithms such as decision trees and neural networks. In addition, they produce highly accurate models and can deal with large datasets [2].
- J48 Decision Tree Algorithm: is an implementation by the WEKA project team of the well-known tree induction algorithm C 4.5 [10]. It follows a greedy iterative approach in building the decision tree. The algorithm partitions the dataset based on the best informative attribute. At each iteration, the attribute with maximum gain ratio

is selected as the splitting attribute. Decision tree classification models have many advantages. They are easy to interpret and are known to have comparable accuracy to other classification models.

Experimental results

The WEKA platform [11] was used in our experiment. It is a well-known data mining software that supports a wide range of data mining algorithms with a friendly user graphical user interface. All models were built using 10-fold cross validation.

Here, we discuss the obtained prediction models. In the J48 decision tree recovery model, shown in Fig. 2, the attribute healthcare personnel appears as the first splitting attribute. This indicates the importance of this information. The model can be interpreted as follows: if the patient is a healthcare personnel, the model predicts recovery. However, if the patient is not a healthcare personnel then the model examines whether he has any pre-existing disease. If the patient suffers from other diseases, the model predicts death, otherwise recovery is predicted. According to this model, healthcare personnel are more likely to survive MERS-CoV infections. This could be due to the vaccinations that healthcare workers are required to get on regular basis.

Fig. 3 shows the stability model using J48 decision tree algorithm. This model shows that the two important attributes for predicting stability are symptomatic and age. The model first checks symptoms, if symptoms exist, then the age of the patient is examined. The status of patients between the ages of 66—87 are predicted as critical. From this model, we conclude that old patients are at high risk of developing MERS-CoV complications.

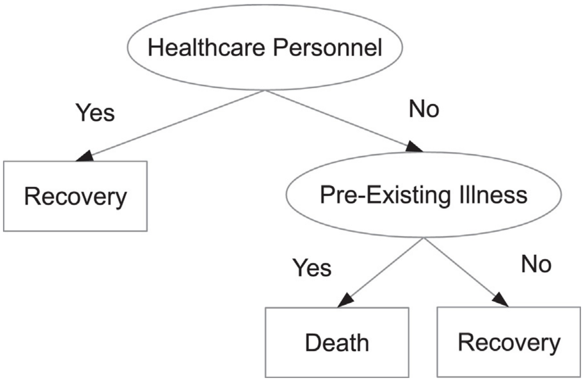


Figure 2 J84 decision tree recovery model.

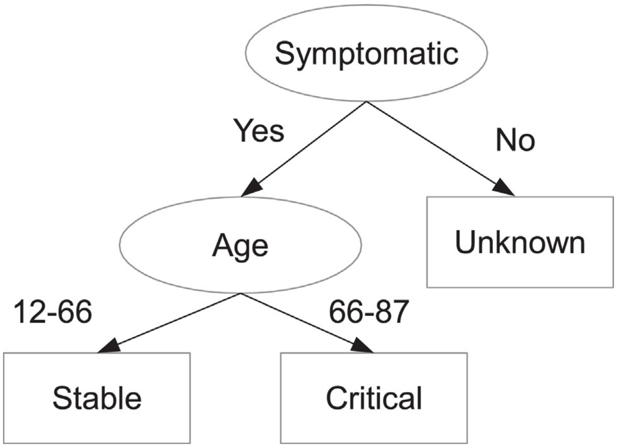


Figure 3 J84 decision tree stability model.

Table 1 Performance evaluation for MERS-CoV recovery models.

Model (year)	Accuracy	Class	Precision	Recall
Naive Bayes	71.58%	Recovery Death	79.4% 66.5%	60.4% 83.4%
J48	68%	Recovery Death	86% 61.3%	45.2% 92.2%

Evaluation

In order to evaluate the accuracy of the obtained models, three performance measures were used: accuracy, precision, and recall. Accuracy refers to the percentage of correctly classified records. Precision is the percentage of records that the model correctly classified as positives out of all positive predictions. Recall measures the true positives recognition rate. These measures are calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{P + N} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

where P is the number of positive records. N is the number of negative records. TP is the number of records that were correctly classified as positive. TN is the number of records that were correctly classified as negative. FN is the number of records that were misclassified as negative.

Tables 1 and 2 summarizes the evaluation measures for the obtained recovery and stability models, respectively. The Naive Bayes recovery

Table 2 Performance evaluation for MERS-CoV stability models.

Model (year)	Accuracy	Class	Precision	Recall
Naive Bayes	53.63%	Stable	56.9%	67.5%
		Critical	41.1%	43.1%
J48	55.69%	Stable	54.9%	89.5%
		Critical	43.9%	15.3%

model performs better in terms of overall accuracy. In addition, it shows high recognition rate for the recovery class. However, the J48 has better recognition rate for class death. As for stability models, the results indicate that J48 has better overall accuracy. However, the recognition rate of class critical is very low. In general, we observe that the performance measures of the recovery models are higher than the stability models.

In this work, we use a real dataset with two classification algorithms known to produce highly accurate models. However, the performance of the all obtained models is not satisfactory for application in real world. The main limitation lies in the size of the training dataset. We believe that there is a need to increase the size of the dataset in order to improve predictions. In addition, more patient information (ex. medical history) can be included.

Conclusion

In this paper, we built several models to predict the stability and recovery of MERS-CoV infections. Our models were built using Naive Bayes and J48 decision tree classification algorithms. The decision tree recovery model indicated that patients who are healthcare personnel are more likely to survive. The age attribute was found to be important in predicting the stability of the patient. Old patients with ages between 66 and 87 are more likely to suffer from critical complications. The performance of all the models was evaluated and compared. In general, the accuracy of the models is between 53.6% and 71.58%. 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.

Funding

No funding sources.

Competing interests

None declared.

Ethical approval

Not required.

Acknowledgment

The project was scientifically supported by King Saud University, Deanship of Scientific research, research chairs and The Research Chair of Health Informatics and Promotion.

References

[1] Coronavirus Website - Ministry of Health. URL http://www.moh.gov.sa/en/CCC/.

[2] Han J, Kamber M. Data Mining: Concepts and Techniques, 3rd Edition The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann; 2011.

[3] Suh SC. Practical Applications of Data Mining. Jones & Bartlett Publishers; 2012.

[4] Ferreira D, Oliveira A, Freitas A. Applying data mining techniques to improve diagnosis in neonatal jaundice. BMC Medical Informatics and Decision Making 2012;12:143.

[5] Venkatalakshmi B, Shivsankar M. Heart disease diagnosis using predictive data mining, International Journal of Innovative Research in Science. Engineering and Technology 2014;3:1873—7.

[6] Majali J, Niranjan R, Phatak V, Tadakhe O. Data mining techniques for diagnosis and prognosis of cancer. IJARCCE 2015;4(3):613—5.

[7] Bellaachia A, Guven E. Predicting breast cancer survivability using data mining techniques, in: Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining; 2006.

[8] Afshar HL, Ahmadi M, Roudbari M, Sadoughi F. Prediction of breast cancer survival through knowledge discovery in databases. Global Journal of Health Science 2015;7(4):392.

[9] Sandhu R, Sood SK, Kaur G. An intelligent system for predicting and preventing MERS-CoV infection outbreak. The Journal of Supecomputing 2015:1—24.

[10] Quinlan R. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann Publishers; 1993.

[11] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor. Newsl 2009;11(1):10—8.