Minimizing False Negatives of Measles Prediction Model: An Experimentation of Feature Selection Based On Domain Knowledge and Random Forest Classifier

Wan Muhamad Taufik Wan Ahmad, Nur Laila Ab Ghani, Sulfeezah Mohd Drus

Abstract: In the context of disease prediction model, false negative error occurs when the patient is wrongly predicted as free from the disease. A prediction model development involves the process of data collection and feature selection which extracts relevant features from the dataset. Two commonly employed feature selection approaches are domain knowledge and data-driven, that suffer from bias towards past or current knowledge when applied alone. In this research, we have studied the development of measles prediction model by incorporating both the domain knowledge and the data-driven approaches, in particular, the Random Forest classifier. The domain expert has earlier on set the important features based upon his prior knowledge of measles for the purpose of minimizing the size of features. Afterward, the attributes became the input in Random Forest classifier and the least important attributes are excluded using the Mean Decrease Gini, in order to experiment its effect on the result. It is found that the removal of several attributes after domain knowledge consultation can provide a good model with less false negative errors.

Keywords: Feature selection, classification, random forest, variable importance, mean decrease gini.

I. INTRODUCTION

In medicine, prediction models are used to predict a patient’s risk of developing a specific disease. False negative refers to the outcome that incorrectly predicts a patient as not having the disease, when the patient does have that disease. A false negative may delay immediate treatment to the patient and increase the risk of fatality (Maxim, Niebo & Utell, 2014). The development of a prediction model starts by collecting data and extracting relevant features of the patient from the data, a process termed as feature selection. Feature selection provides the means to minimize false negative (Balakrishnan et al., 2008) and improve the accuracy of the prediction model (Cai, Wang & Yang, 2018). It works by selecting relevant features of the dataset according to a certain feature selection criterion.

Domain knowledge and data-driven feature selections are the two commonly employed feature selection approaches (Wilcox & Hripcsak, 2003; Groves, 2013; Bochare et al., 2014; Raghu et al., 2017). The domain knowledge approach relies on the expert judgment based on their accumulated years of training and practices. On the other hand, the data-driven approach relies solely on the data and machine learning algorithms to do the feature selection process. However, using either one of the two approaches alone suffers from bias towards current and past knowledge, and do not incorporate all available information to make future predictions (Raghu et al., 2017). It has been pointed out that combining prior domain knowledge as a part of machine learning projects would serve as a complementary to the data-driven approach (Bochare et al., 2014; Islam et al., 2018).

Domain knowledge can be used in the early phase of data preparation to select specific features that are relevant to the prediction task, and the selected features will become the input to the machine learning algorithm being used. Bochare et al. (2014) has incorporated domain knowledge in their breast cancer prediction model and their observation revealed that prediction model generated using both the domain knowledge and data-driven feature selection performed better as compared to the use of data-driven approach alone. Li et al. (2018) has also performed similar experiment for oral disease prediction and deduced that the approach helps in removing irrelevant features and improving prediction accuracy.

Wahet et al. (2018) compared several data-driven feature selection techniques to maximize the accuracy of diabetes and breast cancer prediction model and has suggested for future studies into feature selection using Random Forest classifier. Random Forest classifier is a popular prediction model that can compute the Mean Decrease Gini, which is a measure of feature importance for estimating a prediction outcome (Wang, Yang & Luo, 2016; Jaiswal & Samikannu, 2017). It is widely recognized as a practical method of feature selection due to its capability in estimating the importance of features and providing good predictive performance (Čehovin & Bosnić, 2010; Kawakubo & Yoshida, 2012). It has been implemented to the study on lung cancer prediction (Chicco & Rovelli, 2019), drug toxicity prediction (Hooda, Bawa & Rana, 2018),

Revised Manuscript Received on September 22, 2019

Wan Muhamad Taufik Wan Ahmad, College of Computing and Informatics, Universiti Tenaga Nasional, Jalan IKRAM-UNITEN, 43000 Kajang, Selangor, Malaysia
Nur Laila Ab Ghani, College of Computing and Informatics, Universiti Tenaga Nasional, Jalan IKRAM-UNITEN, 43000 Kajang, Selangor, Malaysia
Sulfeezah Mohd Drus, College of Computing and Informatics, Universiti Tenaga Nasional, Jalan IKRAM-UNITEN, 43000 Kajang, Selangor, Malaysia

DOI: 10.35940/ijeat.A2640.109119

Published By:
Blue Eyes Intelligence Engineering & Sciences Publication

ISSN: 2249 – 8958, Volume-9, Issue-1, October 2019

3411
Minimizing False Negatives of Measles Prediction Model: An Experimentation of Feature Selection Based on Domain Knowledge and Random Forest Classifier

This research is motivated by the alarming increase of measles cases around the world. It has been reported by the World Health Organization that there are more than 30 percent increase of reported measles cases since 2016, with measles death rose by over 20 percent globally in 2017. During a high incidence of measles infection, clinical diagnosis is used to identify measles patients without going for laboratory test confirmation. As a result, a suspected patient can be misclassified as not having measles, when the patient is actually infected. In medicine, it is essential to avoid false negative cases as much as possible, and the challenge in managing measles cases especially during outbreak transmission is to identify measles patient during the earlier stage correctly. The ability to correctly identify possible infected patient provides significant values for planning and action to be taken on those needed. Hence, this research aims to investigate the impact of combining domain knowledge and Random Forest feature selection criterion to the performance of our measles prediction model, focusing on its ability to minimize the false negative error. The research is significant in the sense that we propose a framework of measles prediction model incorporating both domain knowledge and data-driven feature selection method. Important attributes that influence the performance of the prediction model especially on minimizing false negative errors are also identified in this study.

The paper is organized as follows: Section 2 presented the data source and framework of the classification model. Section 3 illustrates the results, before the concluding remarks in Section 4.

II. MATERIALS AND METHODS

Measles cases in Malaysia are used as a case study for this research. Data on measles cases of a one-year period in Malaysia is retrieved in .xlsx format from the database of measles surveillance system, Ministry of Health Malaysia. The data contains 49 attributes and 5,650 records, storing information related to patients’ data, vaccination status, symptoms, diagnosis, laboratory test, result, dates, and the district that the case is being reported.

![Fig. 1 Measles Prediction Model General Framework](image)

Figure 1 illustrates the general framework of our measles prediction model. The collected data is pre-processed by removing attributes with missing values of more than the user-specified threshold. Then, the domain knowledge is acquired in order to understand the attribute definition, determine the outcome attribute and eliminate the irrelevant attributes. Case classification is set as the outcome attribute, which is based on the result of laboratory test, and the total number of attributes is reduced based on the domain expert advice. During feature transformation, records with different values but has a similar meaning are modified, and records with spelling errors are also corrected. For outcome attribute, its records are transformed into binary values as either confirmed cases or discarded cases based on the domain expert suggestion. The transformed dataset is modeled using Random Forest classifier appointing case classification as the outcome, and the variable importance of each feature is measured using Mean Decrease Gini.

The combination of features that can reduce false negative errors investigated by conducting several experiments using different feature sets identified based on the Mean Decrease Gini. The two least important features are removed during the first experiment and then added with other features that have higher importance values. Random Forest modeling is performed to the training dataset of the selected feature set which have been randomly split by 80:20 of training and testing set without having validation set. The model will be tested on the remaining 20 percent of testing data. The subsequent experiments repeated the same process with another feature set. The performance and false negative error in each iteration are recorded for comparison purpose.

III. RESULTS AND DISCUSSION

A total of 19 features is removed during pre-processing due to its large percentage of missing values. Consultation with the domain expert further reduced the features to 15 features. Overall, 15 irrelevant features are
removed by the domain expert that are considered not related in predicting measles lab test result. In this study, domain expert is involved in the pre-modelling phase which will help lessen the scope of the feature selection method of Random Forest. Table 1 shows the description of features selected based on the domain knowledge. The selected features comprise of patient’s personal information, associated symptoms, the state where the cases is reported, type of case, results of rubella and measles test as well as the case classification.

Table 1 also specifies the value of important features based on the Mean Decrease Gini. Each attribute provides different weightage of importance towards model building. Fever, lymphadenopathy, case type, nationality and gender has the smallest value between 1.0 to 4.1. This result suggests that few attributes might be less relevant in developing predictive model and reducing false negative error.

<table>
<thead>
<tr>
<th>Features</th>
<th>Description</th>
<th>Mean Decrease Gini</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age of patient</td>
<td>25.9</td>
</tr>
<tr>
<td>Gender</td>
<td>Gender of patient</td>
<td>4.1</td>
</tr>
<tr>
<td>Nationality</td>
<td>The nationality of the patient</td>
<td>3.5</td>
</tr>
<tr>
<td>Vaccination Status</td>
<td>Vaccination status of patient</td>
<td>26.7</td>
</tr>
<tr>
<td>State</td>
<td>State where the cases is reported</td>
<td>28.1</td>
</tr>
<tr>
<td>Type</td>
<td>Type of cases, either local or imported</td>
<td>3.2</td>
</tr>
<tr>
<td>Outbreak Association</td>
<td>Specify whether the cases reported during an outbreak or not</td>
<td>16.2</td>
</tr>
<tr>
<td>Fever</td>
<td>Specify whether the patient has a fever or not.</td>
<td>1.0</td>
</tr>
<tr>
<td>Cough</td>
<td>Specify whether the patient has a cough or not.</td>
<td>24.7</td>
</tr>
<tr>
<td>Coryza</td>
<td>Specify whether the patient has coryzaor not.</td>
<td>12.4</td>
</tr>
<tr>
<td>Conjunctivitis</td>
<td>Specify whether the patient has conjunctivitis or not.</td>
<td>34.7</td>
</tr>
<tr>
<td>Lymphadenopathy</td>
<td>Specify whether the patient has lymphadenopathy or not.</td>
<td>1.2</td>
</tr>
<tr>
<td>IgM_Rubella</td>
<td>The result of rubella laboratory test</td>
<td>56.8</td>
</tr>
<tr>
<td>1st_IgM_Result</td>
<td>The result of measles laboratory test</td>
<td>412.6</td>
</tr>
<tr>
<td>Case Classification</td>
<td>Indicate whether a patient has measles or not based on laboratory test</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2 shows the performance of the prediction model under three conditions: 1) No application of domain knowledge and Random Forest feature selection, 2) With only the application of domain knowledge feature set, and 3) With the application of both domain knowledge and Random Forest feature selection. The result shows that the removal of fever and lymphadenopathy attributes prior to the model development are able to reduce the false negative error while maintaining high accuracy result. Although the application of feature sets based upon only the domain knowledge produced the highest accuracy, its number of false negative errors is higher than the application of both domain knowledge and Random Forest feature selection. In addition, instead of using 15 attributes, 13 attributes provided better performance for producing less false negative errors. This also will help in the future if the number of records is too high, then the model is easier to be executed due to having less attributes.

<table>
<thead>
<tr>
<th>Features</th>
<th>Without Domain Knowledge and Random Forest Feature Selection</th>
<th>Only Domain Knowledge Features Set</th>
<th>With Domain Knowledge and Random Forest Feature Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>93.5</td>
<td>98.88</td>
<td>98.36</td>
</tr>
<tr>
<td>False Negative Error (%)</td>
<td>0.8</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>Number of False Negative Error</td>
<td>49</td>
<td>32</td>
<td>23</td>
</tr>
<tr>
<td>Number of Attributes Included</td>
<td>30</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>Attributes Excluded based on Feature Importance</td>
<td>n/a</td>
<td>n/a</td>
<td>Fever and Lymphadenopathy</td>
</tr>
</tbody>
</table>

Table. 2 Performance Measurement of the Predictive Model
IV. CONCLUSION

Combination of domain knowledge and random forest feature importance criterion has enabled the production of high accurate model with less false negative error. In that sense, if both domain knowledge and data-driven approach are being considered, the accuracy of the produced result is high and reliable. It can be concluded that the absence of fever and lymphadenopathy attributes help the model in reducing false negative error which is really required to increase confidence for medical practitioners towards relying on machine learning predictive model. It may help to reduce the error of incorrect diagnosis of infected individual. Moreover, having less attributes is preferable as more attributes need more resources such as high processing power computers. Nevertheless, there is a limitation in terms of initial attributes that consisted of a large amount of missing values but considered important in the view of domain expert. Future research can look into ways of minimizing missing values during data collection process.

ACKNOWLEDGMENT

This work was supported by the Ministry of Higher Education, Malaysia under Fundamental Research Grant Scheme(FRGS/1/2016/ICT04/UNITEN/03/1).

REFERENCES


DOI: 10.35940/ijeat.A2640.109119
Published By: Blue Eyes Intelligence Engineering & Sciences Publication