International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

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

Wan MuhamadTaufik Wan Ahmad, NurLailaAb Ghani, SulfeezaMohd 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 developmentof 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 uponhisprior knowledgeon 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 ofseveral 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 doeshavethat 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 (Balakrishnanetal.,2008)and improve the accuracy of the prediction model (Cai, Wang & Yang, 2018).

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Wan MuhamadTaufik Wan Ahmad*, College of Computing and Informatics, Universiti TenagaNasional, Jalan IKRAM-UNITEN, 43000 Kajang, Selangor, Malaysia
NurLailaAb Ghani, College of Computing and Informatics, Universiti TenagaNasional, Jalan IKRAM-UNITEN, 43000 Kajang, Selangor, Malaysia
Sulfeeza Mohd Drus, College of Computing and Informatics, Universiti TenagaNasional, Jalan IKRAM-UNITEN, 43000 Kajang, Selangor, Malaysia

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; Raghuet 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 andmachine 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 allavailable 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 (Bochareetal., 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 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 classifieris 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),

tuberculosis and sarcoidosis classification (Wu, Wang & Wu, 2017), and esophageal cancer prediction (Paul et. al., 2017). 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 negativeerror. The research is significant in the sense that we propose aframework 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 aone-year period in Malaysia is retrieved in *.xlsx* format from the database of measles surveillance system, Ministry of Health Malaysia. The data contains49attributes 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.

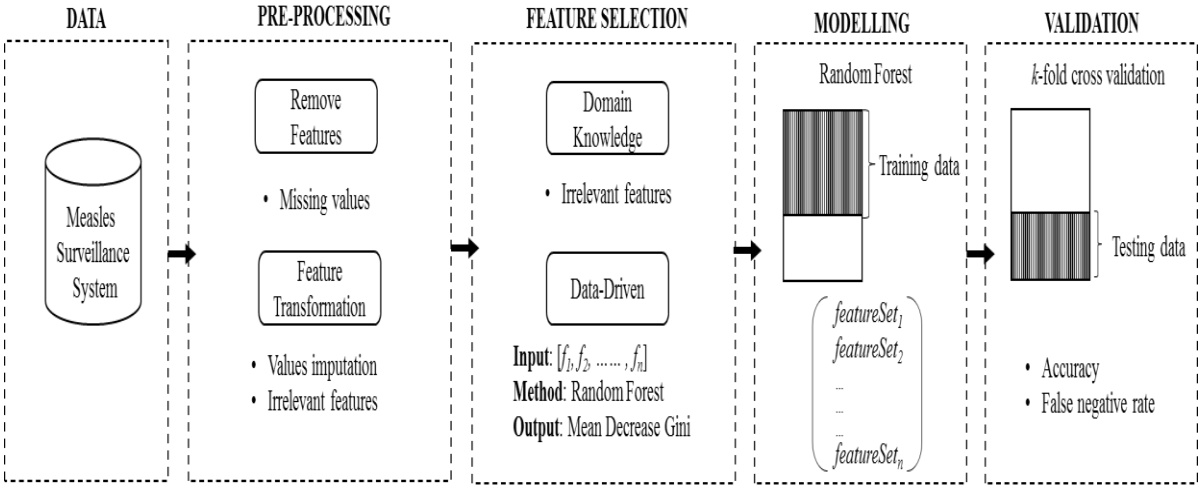

Fig. 1 Measles Prediction Model General Framework

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 theuser-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,andthe total number of attributesis reduced based on the domain expert advice. During feature transformation, records with different values but has a similar meaningare modified,and records with spelling errorsare also corrected. For outcome attribute, its records are transformed into binary values as either confirmed measles 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 negativeerroris investigated by conducting severalexperimentsusing different feature set identified based on the Mean DecreaseGini. The two least important featuresare removed during the first experiment and then added with otherfeatures that have higher importance values.Random Forest modelingis performed to the training dataof the selected feature setwhich 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 processwith another feature set. The performance and false negativeerror in each iterationare recorded for comparison purpose.

III.RESULTS AND DISCUSSION

A total of 19 featuresis removed during pre-processing due to its large percentage of missing values. Consultation with the domain expert furtherreduced the features to 15 features.

Overall, 15 irrelevants 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, thestate where the caseis 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. 1 Description of Selected Features and Its Mean Decrease Gini Value

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

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. 2 Performance Measurement of the Predictive Model

	Without Domain Knowledge and Random Forest FeatureSelection	**Only Domain Knowledge Features Set**	**With Domain Knowledge and Random Forest Feature Selection**
Accuracy (%)	95.5	98.88	98.36
False Negative Error (%)	0.8	0.5	0.4
Number of False Negative Error	49	32	23
Number of Attributes Included	30	15	13
Attributes Excluded based on Feature Importance	n/a	n/a	Fever and Lymphadenopathy

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 practioners towards relying on machine learning predictive model. It may helps to reduce the error of incorrect diagnosis of infected individual. Moreover, having less attributes is preferrable as more attributes need more resources such as high processing power computers. Nevertheless, there is a limitation in termsof initial attributes that consisted ofa 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

1. Balakrishnan, S., Narayanaswamy, R., Savarimuthu, N., & Samikannu, R. (2008, October). SVM ranking with backward search for feature selection in type II diabetes databases. In 2008 IEEE International Conference on Systems, Man and Cybernetics (pp. 2628-2633). IEEE.
2. Bochare, A., Gangopadhyay, A., Yesha, Y., Joshi, A., Yesha, Y., Brady, M., ... & Rishe, N. (2014). Integrating domain knowledge in supervised machine learning to assess the risk of breast cancer. International Journal of Medical Engineering and Informatics, 6(2), 87-99.
3. Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70-79.
4. Čehovin, L., & Bosnić, Z. (2010). Empirical evaluation of feature selection methods in classification. Intelligent data analysis, 14(3), 265-281.
5. Chicco, D., & Rovelli, C. (2019). Computational prediction of diagnosis and feature selection on mesothelioma patient health records. PloS one, 14(1), e0208737.
6. Groves, W. (2013, June). Using domain knowledge to systematically guide feature selection. In Twenty-Third International Joint Conference on Artificial Intelligence.
7. Hooda, N., Bawa, S., & Rana, P. S. (2018). B2FSE framework for high dimensional imbalanced data: A case study for drug toxicity prediction. Neurocomputing, 276, 31-41.
8. Islam, M., Hasan, M., Wang, X., & Germack, H. (2018, June). A systematic review on healthcare analytics: Application and theoretical perspective of data mining. In *Healthcare* (Vol. 6, No. 2, p. 54). Multidisciplinary Digital Publishing Institute.
9. Jaiswal, J. K., & Samikannu, R. (2017, February). Application of random forest algorithm on feature subset selection and classification and regression. In 2017 World Congress on Computing and Communication Technologies (WCCCT) (pp. 65-68). IEEE.
10. Kawakubo, H., & Yoshida, H. (2012, July). Rapid feature selection based on random forests for high-dimensional data. In Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA) (p. 1). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).
11. Li, G., Zhang, S., Liang, J., Cao, Z., & Guo, C. (2018, December). Augmenting Embedding with Domain Knowledge for Oral Disease Diagnosis Prediction. In *International Conference on Smart Computing and Communication* (pp. 236-250). Springer, Cham.
12. Maxim, L. D., Niebo, R., & Utell, M. J. (2014). Screening tests: a review with examples. Inhalation toxicology, 26(13), 811-828.
13. Paul, D., Su, R., Romain, M., Sébastien, V., Pierre, V., & Isabelle, G. (2017). Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. Computerized Medical Imaging and Graphics, 60, 42-49.
14. Raghu, V. K., Ge, X., Chrysanthis, P. K., & Benos, P. V. (2017, April). Integrated theory-and data-driven feature selection in gene expression data analysis. In 2017 IEEE 33rd International Conference on Data Engineering (ICDE) (pp. 1525-1532). IEEE.
15. Wah, Y. B., Ibrahim, N., Hamid, H. A., Abdul-Rahman, S., & Fong, S. (2018). Feature Selection Methods: Case of Filter and Wrapper Approaches for Maximising Classification Accuracy. Pertanika Journal of Science & Technology, 26(1).
16. Wang, H., Yang, F., & Luo, Z. (2016). An experimental study of the intrinsic stability of random forest variable importance measures. BMC bioinformatics, 17(1), 60.
17. Wilcox, A. B., & Hripcsak, G. (2003). The role of domain knowledge in automating medical text report classification. Journal of the American Medical Informatics Association, 10(4), 330-338.
18. Wu, Y., Wang, H., & Wu, F. (2017, October). Automatic classification of pulmonary tuberculosis and sarcoidosis based on random forest. In 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) (pp. 1-5). IEEE.