Semi Supervised Learning to Classify Drug Resistant Tuberculosis

Prabu Setyaji¹ and Preethi Subramanian²

¹Asia Pacific University (APU), Kuala Lumpur, Malaysia.

ABSTRACT

Background: Health is one of the vital factors for human survival and continuous efforts are focused on research in this domain. Spreading of tuberculosis increases from year to year and is a life threat that exists since antiquity.

Problem: The problem gets complicated with other viruses like HIV and drug-resistant tuberculosis. World Health Organization has published data about drug-resistant tuberculosis to analyze the problems faced.

Objective: This paper focuses on applying a semi-supervised learning model to prescribe recommendations based on data analytics. Three models such as the Decision Tree, Gradient Boosting and Neural Network are trained to predict the clusters. Gradient Boosting can perform the best with the lowest misclassification rate and the majority cluster is identified based on its impact and population.

Conclusion: The outcome of this analysis can provide recommendations to the health domain to reduce the spread of diseases like tuberculosis and also enhance the preparedness in terms of drug production.

Key Words: Semi Supervised Learning, Gradient boosting, Prescriptive analytics

INTRODUCTION

World Health Organization (WHO) is the global governing body which oversees health issues and related concerns across several countries and continents. One of the major health threats is Tuberculosis (TB) as there are 10 million active cases with 1.5 million death cases. Moreover, TB cases increase undiagnosed each year at a rate of 3 million¹. TB is one of the oldest disease caused by Mycobacterium Tuberculosis bacteria that infected human lung alveoli (respiration tube). It is known to have killed an estimated 2 million humans every year according². It can be cured by vaccine and antibiotics, however, the problem of TB persists and one of the main issues is drug resistance for the TB patients. WHO³ stated that drug-resistant issue came from the misuse of antibiotics in chemotherapy. A survey³ on these drug-resistant patients is not conclusive with inaccurate responses. Current data and analysis published by WHO cannot estimate the future of drug-resistant TB cases and this leads to the inability in predicting the estimate of drug production.

At the moment WHO merely has drug-resistant TB data and analysis up until 2017⁴. It has been stated that 56% of the patients have not been cured and referring to the data that the organization published the number of drug-resistant TB patients increase from 2017 to 2018 and thereafter. Mezwa et al.⁵ added that drug resistance can exaggerate into Multi-Drug Resistant (MDR) cases where that patient is immune to more than one antibiotics such as Rifampicin and Isoniazid. Khan Academy⁶ has stated that these two drugs are the most important first-line drug for TB medicine as second-line drugs are only for emergency and they are also expensive. Extensive Drug Resistant (XDR) is far worse than second-line drugs would not be effective and the condition implies that new drugs will be needed. The number of deaths and drug-resistant cases needs to be predicted so that future estimates can help in drug production planning. The paper focuses on applying data analytical methods to predict future cases and perform prescriptive analytics.
MATERIALS AND METHODS

This research utilizes the dataset from the WHO website and the dataset contains data for drug-resistant Tuberculosis. The data exists for over 200 countries for the year 2018. CRISP-DM was adopted to manage the task accordingly.

Cluster Identification and Profiling

The dataset is given by WHO appears crude and required a lot of data wrangling steps. Due to the nature of the dataset, the clustering division is not equal as most of the observations belong to cluster number 3. Careful consideration has been provided to select the clustering algorithms and the number of clusters. However, an uneven sized cluster was concluded as the best result as shown in Figure 1.

Then Table 1 shows the frequency of each cluster that helps to clarify the characteristics of each cluster. It was found that Cluster 2 has the least number of observations and this cluster may not have a high impact. The cluster distance table can be seen in Table 2.

Table 1: Identification of Clusters

<table>
<thead>
<tr>
<th>Clustering Criterion</th>
<th>Maximum Relative Change in Cluster Seeds</th>
<th>Segment Id</th>
<th>Frequency of Cluster</th>
<th>Root-Mean-Square Standard Deviation</th>
<th>Maximum Distance from Cluster Seed</th>
<th>Nearest Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.238542</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>0.754512</td>
<td>25.35754</td>
<td>2</td>
</tr>
<tr>
<td>0.238542</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>12.18368</td>
<td>176.98815</td>
<td>2</td>
</tr>
<tr>
<td>0.238542</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>0.98815</td>
<td>8.91836</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2: Inter-Cluster Distances

The modelling procedure continues to profile the segments obtained, where the result of each cluster can be displayed to emphasize the characteristic of each cluster and which variables have a high impact as shown in the Appendix in Figure 7. Although Cluster 3 and Cluster 1 shares the same important variables, the segment profile shows that Cluster 3 holds 176 observations which are the majority. However, it does not signify that Cluster 1 is less important because it identifies DsrRlt_Rr new and DsrRlt_Rr ret variables that represent people that have drug-resistant TB. This detail explains the drug capability of handling patients. On the other hand, Cluster 2 does not stand a chance to be focused because of the negligible number of observations. Overall, Cluster 3 is the most significant as it contains an important variable like pulmlabconf rets which identifies the people that relapse or still sick from the previous treatment. This will help to regulate pulmonary TB and can be noted as a finding from this analysis.

Moving on in the analysis, data wrangling had been carried out with missing value imputation being performed on the complete dataset using average/mode values. The data was partitioned before feeding into the predictive models. Models such as Decision Tree, Gradient Boosting and Neural Network were designed to predict the target. The models predicted the segment variable or cluster-id created in the earlier phase.

Critical Interpretation of the models

In this section, the focus will be on interpretation of the result in term of action that can be conducted to assess the business issue of the organization.

Decision Tree

Decision Tree produces a satisfactory result as it has a low misclassification rate on training but, quite high on validation that represents the main measure of the classification model in predicting the value as Figure 2 shows the plot of misclassification rate of the model.
Figure 3 details the decision tree and it shows that the model can predict cluster 3 with high confidence and cluster 1 as well. One of the reasons for the higher prediction of Cluster 3 is because of the number of observations that belong to that cluster. Results of the decision tree show the variables that affect model accuracy and RR_ret is the number of patients that have been previously treated for drug-resistant Tuberculosis and still ongoing/relapse. pulmlabconf ret is the number of pulmonary Tuberculosis patients who have been treated previously.

Figure 3: Decision Tree Result.

Both the important variables support the emphasis of preparing drugs for drug-resistant Tuberculosis.

Gradient Boosting

Among other models, Gradient Boosting has performed the best as it has produced a simple tree structure with high accuracy and also maintains a low misclassification rate for both train and validation as shown in Figure 4.

Figure 4: Gradient Boosting Misclassification Rate.

The tree produced by the Gradient Boosting has more nodes compare to the Decision Tree. More variables are being considered and impacted by the model. Pulm lab conf ret still maintain their importance but with the addition of other attributes as in Figure 5. The first variable displayed as the filter is DstRltRr Ret that represent the number of patients tested with first-line drugs and having resistance to Rifampicin. PulmLabconf New is the new patients identified having pulmonary Tuberculosis.

Figure 5: Gradient Boosting Result.

These two variables seem to have a huge impact on the prediction as a new identified Tuberculosis patient could lead to drug-resistant cases. Identifying patient resistance to one of the first-line drugs is vital to study the need for the effectivity of Rifampicin that acts as the first-line drug for Tuberculosis patients.

Neural Network

Neural Network is the third model that is tested to estimate the target and it surprisingly has the highest misclassification rate on validation data and it has a different behaviour compared to the rest. Neural Network chosesr_hivpos as the only option in the Neural Network but it never appeared according to the other techniques tried out. Also, Neural Network only predicts dominant clusters such as cluster 1 and 3. Interpretation of the surrogate tree result is shown in Figure 6 below.

Despite the differences in the estimation, this model can still the clusters very well, however, the variables that are employed are varied and it is more focused on HIV despite its connection to drug-resistant.
Interpretation Summary

Overall summary of the interpretation will be discussed in this section and Table 3 states a comparison amongst models to point out which one is the best model to be used to solve the issues based on misclassification rate.

Table 3: Fit Statistics

Refer to Table 3, Gradient Boosting has the best misclassification rate followed by Decision Tree Then Neural Network with the highest misclassification. With this result, Gradient Boosting is the perfect fit model to solve the issue that is being discussed to predict drug-resistant Tuberculosis based on the main number of Tuberculosis patients from previous and new identifications. It was also found that the number of patients tested and immune to Rifampicin regardless of Isoniazid is one of the important criteria to predict the cluster. It is evidenced that Cluster 3 is the most important of them all due to its vast percentage and also since most of the variables impacted are the necessary ones to prescribe the drug-resistant (RR) value.

CONCLUSION AND RECOMMENDATIONS

Selected data analytical models were employed to classify the drug-resistant tuberculosis data based on the clusters identified. Gradient Boosting model was identified to have the best prediction performance and this model found the three major variables that could explain the probable outcomes. The model can identify drug-resistant TB patients and also previous relapse patients with complete details. With these hidden patterns uncovered in data, the health authorities would be able to predict the future of drug-resistant TB by studying the historical data. According to the data studied from WHO, the observations were divided into 3 clusters and Cluster 3 hold the major decision in predicting the possible incoming drug-resistant TB. With the presence of Cluster 3 cases, the analysis also provides details on the countries or regions that are more vulnerable to the drug-resistant TB. The analysis also reveals the type of drug-resistant TB, the effectiveness of the lines of injectable drugs. Further improvement can still be applied to the model as the Gradient Boosting has utilized all default parametrizations in this analysis. A similar analysis can also aid the health authorities to be better prepared in cases of pandemics, drug production and improved medical attention by prescribing the right drugs based on the variations observed.

ACKNOWLEDGEMENTS

The authors also wish to express gratitude to the management of Asia Pacific University of Technology & Innovation (APU) for their support.

Conflict of Interest

The authors involved in the current study does not declare any competing conflict of interest.

Funding and Sponsorship

No fund or sponsorship in any form was obtained from any organization for carrying out this research work.

REFERENCES


APPENDIX

Figure 7: Segment Profile and Characteristics of Clusters.