Analysis of Classification Techniques for Prediction of Tuberculosis Defaulters

Brian Mc George
University of Cape Town
Cape Town, South Africa
mcgbri004@myuct.ac.za

ABSTRACT

1. INTRODUCTION

In 2013 over 210 000 patients defaulted from Tuberculosis (TB) treatment worldwide [4]. The rate of default in the Americas is the highest at 8% with Africa at 5% [4]. The consequences of defaulting TB treatment include: increased drug resistance, increased health system costs [2, 3], higher risk of mortality, continued risk of transmitting the disease to others [2] and increased rate of recurrent disease [1]. The spread of TB can be reduced if the individuals who have a high risk of defaulting can be predicted. This will also reduce health system costs.

The field of credit scoring in the financial space aims to determine if a financial institution should provide credit to an individual. This area has been well researched. This paper aims to determine if classification techniques that have been evaluated for the credit scoring problem will show similar results for predicting TB defaulters. There are notable similarities in these problems which could make them comparable. Both problems typically have a labelled dataset consisting of both nominal and numerical data as well as the occurrence of missing data [citation needed]. However, TB datasets are more prone to missing data as well as inaccuracies due to the nature of the data collection [citation needed]. The selected classification techniques are evaluated against real-world treatment default datasets and financial datasets. The paper will evaluate how the techniques differ across the datasets. If the relative results are similar then future credit scoring research could be applicable to treatment default prediction too.

1. Discuss datasets used and possibly expansion of TB issues in those specific counties (Peru and Malawi)
2. Outline briefly how the datasets are used and that each technique is also benchmarked against the well known Australian and German financial to determine how applicable credit scoring research is to TB default prediction for the two TB datasets.

DOI: n/a

3. Link TB classification to credit scoring and outline notable similarities and differences
4. Summarise overall paper

2. RELATED WORK

1. Summarised version of original literature review except the parts using temporal aspects
2. Summarise what has been done determining TB predictors (which is a different aim)
3. Summarise what has been done in the credit scoring financial space as a lot of research has been done in this field.

3. METHOD

1. Outline each dataset fully: number of nominal and numerical fields and number of entries. Also outline how balanced each dataset is.
2. Outline testing procedure and optimisation strategy for each classification technique. To optimise each classification technique, a grid search is conducted across a reasonable parameter space. For each parameter set, the results are averaged over 3 runs, this is done because of the stochastic nature of initialising the training of each classifier. To ensure consistent results for each run, the same fold allocations are used.
3. Brief overview of each classification technique and possibly give a reason why this technique may work well for our application

4. RESULTS

1. Tables with true positive, true negative, false positive and false negative rate for each classifier on each dataset
2. ROC curves which can be used to determine true positive for an acceptable amount of false positives

5. DISCUSSION

1. Discussion on best classification technique
2. Outline similarities and differences in results between the two TB datasets and between TB and financial datasets and determine if they are similar enough that results for the German and Australian credit scoring datasets could be used for the two TB datasets.

6. FUTURE WORK

1. Likely something along the lines of utilising more temporal based data so that classification is not just done at registration but also at each check-up for example. Future work may also be the testing of more classification techniques as well as datasets from other parts of the world.

7. CONCLUSIONS

8. REFERENCES

[1] U. M. Jha, S. Satyanarayana, P. K. Dewan, S. Chadha, F. Wares, S. Sahu, D. Gupta, and L. S. Chauhan. Risk factors for treatment default among re-treatment tuberculosis patients in india, 2006. *PLoS ONE*, 5(1):1–7, January 2010.

[2] B. Lackey, C. Seas, P. Van der Stuyft, and L. Otero. Patient characteristics associated with tuberculosis treatment default: A cohort study in a high-incidence area of lima, peru. *PLoS ONE*, 10(6):1–11, 2015.

[3] B. Muture, M. N. Keraka, P. K. Kimuu, E. W. Kabiru, V. O. Ombeka, and F. Oguya. Factors associated with default from treatment among tuberculosis patients in nairobi province, kenya: A case control study. *BMC Public Health*, 11(1):696–105, September 2011.

[4] World Health Organisation. Global tuberculosis report 2015.