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 [8]. The rate of default in the Americas is the highest at 8% with Africa at 5% [8]. The consequences of defaulting TB treatment include: increased drug resistance, increased health system costs [5, 6], higher risk of mortality, continued risk of transmitting the disease to others [5] and increased rate of recurrent disease [3]. 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. BACKGROUND

2.1 Definition of a defaulter

The definition of a defaulter depends on its context. TB literature typically uses the World Health Organisation (WHO) definition that a defaulter is a person whose treatment has been disrupted for two or more consecutive months [1, 2, 3, 4, 6, 8].

2.2 Determining predictors of TB default

There have been many studies which focus on determining the factors associated with TB default but few have used machine learning techniques to predict treatment defaulters. Table 1 contains an overview of a selection of publications on determining the factors associated with TB default. The majority of techniques use a form of logistic regression ti determine the association.

The datasets used by the publications contain different features. Age and gender are common throughout the datasets. History of past default is available for all datasets except for Shargie *et al.* [7]. Lackey *et al.* [5] only picked individuals who did not have a history of past default. Jittimanee [4] [28] was the only publication with the feature that did not find it to be significant to the 95% confidence level. However, it did have an odds ratio of 2.19 and a p-value of 0.12. It can therefore be deduced that a history of past default has a strong correlation to default. Two out of three publications with the alcohol abuse feature available, found it to be significant. Three of the four publications with side effects, as a feature found it was significant. Shargie *et al.* [7] and Jittimanee *et al.* [4] measured distance and time to treatment site respectively. It can be reasoned that the aforementioned feature's significance will generalise to other datasets since they were found to be significant in the majority of the publications. Other significant features such as illegal drug use, use of herbal medication, daily jobs, history of lung cancer and history of liver disease only appeared once in the datasets. It cannot be discerned if the significance is generalisable or specific to the dataset. The identification of the same features as significant is fairly consistent for the publications that have those features in their dataset.

2.3 Predicting defaulters in financial institutions

Table 1: Overview of publications on predictors of TB treatment defaulters

Publication	Sample Size	Key factors identified*	Evaluation
Chan-Yeung *et al.* [1]	1768 non-defaulters 442 defaulters.	History of default, history of lung cancer, liver disease and male patients	Multiple logistic regression is used to determine what factors are associated with default.
Jha *et al.* [3]	1189 non-defaulters 1141 defaulters.	Male patients, patients directly-observed treatment at public facilities, previous treatment outside India's Revised National Tuberculosis Control Programme and history of previous default	The chi-square test or Fisher's exact test (if there were less than 10 observations) was used to test the differences between defaulters and non-defaulters. Bivariate analysis was calculated on the features. Multivariate logistic regression using pre-selected features based on previous studies.
Jittimanee *et al.* [4]	106 non-defaulters 54 defaulters.	Jobs where one is only paid if one is at work that day, severe medication side-effects and time to travel to clinic.	Patients were interviewed and completed a questionnaire to obtain the information. Hierarchical logistic regression was carried out to assess the variable's relation to default.
Lackey *et al.* [5]	1106 non-defaulters 127 defaulters.	Has used illegal drugs, has multidrug-resistant TB, has not been tested for HIV, drinks alcohol at least once a week, underweight or has not completed secondary education.	Patients were interviewed to obtain the information. Bivariate analysis using Chi-square tests and odds ratios with 95% confidence intervals. Multivariate logistic regression is used with a backward fitting algorithm to determine if a variable is associated with default.
Muture *et al.* [6]	5659 non-defaulters 945 defaulters.	Inadequate knowledge on TB, herbal medication use, low income, alcohol abuse, previous default, suffering from HIV and male patients were determined through analysis to be associated with default.	Two-tailed χ^2 tests and Fisher exact tests (if a cell has less than 5 values) to assess categorical information. Odds ratio tests were used to measure association between features with a 95% confidence interval.
Shargie *et al.* [7]	310 non-defaulters 81 defaulters.	Distance from home to treatment site, age greater than 25 and the need to use public transport to get to treatment site.	Continuous features were analysed using sample t-test χ^2 tests and hazard ratios with 95% confidence intervals. The effect of each factor was assessed using Cox regression model with a backwards fitting algorithm.

*Ordered in descending order of significance based on odds ratio and hazard ratio at a confidence interval of at least 95%.

Table 2: Predicting financial defaulters using SVM¶

Publication	Accuracy	Type I Error	Type II Error
Huang *et al.* [?]	**79.87%**†	n/a	n/a
Li *et al.* [?]	**84.83%**	10-20%*	10-20%*
Luo *et al.* [?]	77.06% (MySVM), 82.41% (SVM-GA)†	n/a	n/a
Huang *et al.* [?]	82.41% (SVM-GA)†	n/a	n/a
Danenas *et al.* [?]	**94.41%** (Linear SVM), **92.37%** (PSO-LinSVM)†	n/a‡	n/a‡

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] M. Chan-Yeung, K. Noertjojo, C. Leung, S. Chan, and C. Tam. Prevalence and predictors of default from tuberculosis treatment in hong kong. *Hong Kong Medical Journal*, 9(4):263–270, 2003.

[2] I. Cherkaoui, R. Sabouni, I. Ghali, D. Kizub, A. C. Billioux, K. Bennani, J. E. Bourkadi, A. Benmamoun, O. Lahlou, R. E. Aouad, and K. E. Dooley. Treatment default amongst patients with tuberculosis in urban morocco: Predicting and explaining default and post-default sputum smear and drug susceptibility results. *PLoS ONE*, 9(4):1–9, April 2014.

[3] 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.

[4] S. X. Jittimanee, E. A. Madigan, S. Jittimanee, and C. Nontasood. Treatment default among urban tuberculosis patients, thailand. *International Journal of Nursing Practice*, 13(6):354–362, 2007.

[5] 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.

[6] 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.

[7] E. B. Shargie and B. Lindtjorn. Determinants of treatment adherence among smear-positive pulmonary tuberculosis patients in southern ethiopia. *PLoS Med*, 4(2):1–8, February 2007.

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