UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Department of Computer Science

Computer Science Honours
Final Paper
2016

Title: An Analysis of Classification Techniques for the Prediction of Tuberculosis Defaulters and Community Health Worker Attrition

Author: Brian Mc George

Project abbreviation: ML4CHW

Supervisor: Dr Brian DeRenzi

Category	**Min**	**Max**	**Chosen**
Requirement Analysis and Design	0	20	0
Theoretical Analysis	0	25	0
Experiment Design and Execution	0	20	18
System Development and Implementation	0	15	10
Results, Findings and Conclusion	10	20	19
Aim Formulation and Background Work	10	15	13
Quality of Paper Writing and Presentation	10		10
Quality of Deliverables	10		10
Overall General Project Evaluation	0	10	0
Total marks			80

An Analysis of Classification Techniques for the Prediction of Tuberculosis Defaulters and Community Health Worker Attrition

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

ABSTRACT

As the use of electronic data capturing for community health projects becomes more widespread, the availability of this type of data for research use is becoming more prevalent. In this paper we outline various issues that may arise when attempting to apply classification techniques to predict minority class events and apply several strategies to counteract them. We focus on the prediction of tuberculosis defaulters and the attrition of community health workers. We compare 13 classification techniques, 13 data balancing techniques and 6 feature selection techniques to assess what combination produces the best classification results. We find that the use of a data balancing technique greatly improves balanced accuracy and H-measure on imbalanced datasets. We recommend the use of logistic regression, artificial neural networks, random forest and Bernoulli naive Bayes for different use-cases. Adaptive synthetic sampling is recommended for highly imbalanced datasets. Random forest is recommended for identifying the most relevant features. SVM (linear) is recommended for removing redundant features that slow training times, however, our testing suggests it is only worthwhile for large datasets with many features.

CCS Concepts

•**Applied computing** → **Health informatics;**

Keywords

Classification; tuberculosis default; health worker attrition

1. INTRODUCTION

Classification techniques can be used to flag individuals who have a high probability of certain events occurring. In this paper we outline various issues that may arise when attempting to apply classification techniques to predict minority class events and apply several strategies to counteract them. We focus on the prediction of Tuberculosis (TB) defaulters and the attrition of community health workers (CHW).

TB is a global problem and communities often have limited resources to track and follow-up with patients. A patient is considered to have defaulted if their treatment is interrupted for longer than a set duration; typically two consecutive months for TB treatment [63]. In 2013 over 210 000 patients defaulted from TB treatment worldwide [63]. The rate of default in the Americas is the highest at 8% with Africa at 5% [63]. The consequences of defaulting TB treatment include: increased drug resistance, increased health system costs [35, 47], higher risk of mortality, continued risk of transmitting the disease to others [35] and increased rate of recurrent disease [30]. If we can better manage resources by predicting those patients with a high risk of default then the spread of TB can be reduced. This would also reduce health system costs.

Attrition is the loss of workers through resignation or abandonment of the work. If we can flag workers who are likely to quit quickly enough then various intervention measures could be implemented in order to retain those workers. This could reduce costs because new workers would not have to be trained.

The aim of this paper is to inform the reader on issues that can occur when building classification models from imbalanced datasets. Bias can occur in the models since most classifiers expect equal weighting of samples for each class. This can lead to poor classification for the minority class, which is often the class of interest. The second aim of this paper is to propose recommendations on classification techniques, data balancing techniques and strategies to improve the overall classification for this problem domain. We examine a number of different data balancing techniques which either over-sample the minority class, under-sample the majority class or provide a combination of an over-sampling and under-sampling technique. We provide a large scale comparison of different classification techniques in order to determine which technique is best suited to these types of problems. As part of our evaluation of each classification technique, we examine how well each classifier works out-of-the-box compared when they are tuned by searching a grid of parameters. Furthermore, we apply a number of feature selection strategies to determine which strategy is best suited to remove redundant and noisy variables.

In addition to our TB default dataset and attrition dataset, we include two real-world credit scoring datasets. The field of credit scoring in the financial space aims to determine if a financial institution should provide credit to an individual. This is a well researched binary classification problem. We included these as all the problems try to flag high-risk individuals and it will allow us to get a better understanding of how the results compare and generalise.

From our findings we recommend the use of logistic regression (LR), artificial neural networks (ANN), random forest (RF) and Bernoulli naive Bayes for different use-cases. All the classifiers produced good classification results. LR is recommended when it is important to be able to understand how the model is producing the classification. ANN and RF are recommended when raw classification performance is the main concern. Bernoulli NB scales well with dataset size and requires very little tuning. Adaptive synthetic sampling (ADASYN) is recommended for highly imbalanced datasets. RF is recommended for identifying the most relevant fea-

This paper was prepared as part of the final project for a Bachelors of Science Honours in Computer Science from the University of Cape Town in 2016. Additional supplementary material is available at: https://people.cs.uct.ac.za/~mcgbri004/honours

tures. Support vector machine (SVM) with linear kernel is recommended for removing redundant features that only slows the classification, however, our results suggest that the reduction in training time is only useful for large datasets with many features.

2. BACKGROUND

This section aims to provide an overview of all the techniques and metrics used in this paper.

2.1 Classification techniques

We selected a variety of classification techniques to benchmark. We include a selection of well known techniques, ensemble techniques and newer techniques that have shown promising results in other studies. Our choices are based partly on our review of classification techniques used in the credit scoring field discussed in Section 3.2.

2.1.1 Support Vector Machines

SVM maps input vectors into a high-dimensional feature space with the use of a kernel function [14]. The kernel function determines if the mapping is linear or non-linear [43]. The radial basis function (RBF) and polynomial kernels perform non-linear mappings. A model is created by searching the high-dimensional feature space to acquire an optimal hyperplane that separates the space with the maximum distance between each class [14].

2.1.2 Artificial Neural Network

An ANN is comprised of an input layer, k number of hidden layers and an output layer. Neurons in each layer are connected to the neurons in the next layer with a numeric weight between each pair of connected neurons. An activation function defines if a neuron will fire and introduces non linearity into the model [62]. It has been proven that an ANN is able to approximate any continuous function if it has at least one hidden node and an activation function that is both bounded to some range and non-constant [23].

Extreme learning machines (ELM) are an alternative approach to the conventional ANN. Huang *el al.* [28] proved that the input and hidden layer weights can be randomly assigned if the activation function in the hidden layer is infinitely differentiable. By randomly assigning these weights, the weights for the output nodes can be determined analytically [28]. This allows ELMs to be trained orders of magnitude faster than a back-propagated ANN [27, 28]. ELMs have been shown to produce better classification than ANN and SVM with RBF kernel [27, 28].

2.1.3 Logistic Regression

LR [46, 53] is a technique that models the chance of an outcome based on the input features. Since chance is a ratio, the logarithm of the chance is modelled instead [53]: $\log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m$. p represents the probability of an event. β_0 represents the value of the criterion when the predictor is equal to 0. $\beta_1, ..., \beta_m$ are the regression coefficients associated with the $x_1, ..., x_m$ input features.

2.1.4 k-Nearest Neighbours

The k-Nearest Neighbours (kNN) algorithm determines the output classification by examining the k nearest training examples in the feature space [33]. An input is classified by the majority vote of its neighbours.

2.1.5 Ensembles

An ensemble classifier is one which typically makes use of the aggregation of multiple classifiers. A single decision tree (DT) can be used for classification by branching on conjunctions of features and having the leaves represent the output class. However, DT typically provides relatively poor classification accuracy [57].

RF is a technique that fits a number of decision trees on random samples with replacement of the dataset. For each tree, n features are randomly selected and the tree is grown [9]. The output classification is decided by the majority vote of each tree [9].

AdaBoost [18] uses a weighted sum of multiple classifiers in order to determine the output. The classifier is constructed in an iterative fashion. At each iteration a pool of classifiers are considered and one classifier is added to the committee of classifiers. Input which is still misclassified is assigned a higher weight at each iteration which guides the classifier selection process [7, 52].

2.1.6 Naive Bayes

Naive Bayes (NB) assumes that each input feature is independent of each other which simplifies the algorithm [40, 51]. NB uses conditional probability to classify new inputs [40]. Using Bayes theorem, the following equation is derived: $p(C_i|\mathbf{x}) = \frac{p(C_i)\times p(x_1|C_i)\times ...\times p(x_n|C_i)}{p(\mathbf{x})}$ with input $\mathbf{x} = (x_1, ..., x_n)$ and class label C_i [40, 51]. To classify an input, we calculate the probability for each class and pick the class with the highest probability. [40]. The event model of NB describes the assumed distribution of features. The Gaussian event model assumes that $p(x_i|C_i)$ follows a Gaussian distribution [32]. This allows support of continuous x_i values [32]. The multivariate Bernoulli event model assumes that features are independent Boolean values [45].

2.1.7 Clustering-launched classification

Clustering-launched classification (CLC) first clusters the data into groups using a diverse hierarchical k-means algorithm [43]. The clusters are divided into positive subsets and negative subsets [43]. Support vectors are then used to separate the positive and negative subsets for each cluster [43]. Redundant or repeated support vectors are removed thereafter [43].

2.2 Data Balancing algorithms

This section examines the different data balancing approaches and their respective algorithms. Our problem space often has much fewer samples in the class of interest than the other class. Most classifiers expect an equal number of samples per class otherwise they will bias heavily towards the majority class. We include an analysis of several data balancing algorithms to ensure our classifiers are able to detect both the positive and negative class as best as possible. We chose our balancing techniques to facilitate an evaluation of both over-sampling and under-sampling techniques. Some of the techniques use simple approaches and others use sophisticated algorithms.

2.2.1 Over-sampling

Random over-sampling (ROS) is a technique that randomly replicates examples in the minority class [5]. However, this technique can increase the likelihood of over-fitting since exact copies are made from the minority class [5].

The Synthetic minority over-sampling technique (SMOTE) [11] forms new minority samples by interpolating along the line segment on some or all of the k nearest minority class neighbours of a minority example. This approach attempts to alleviate the over-fitting that can occur from using random over-sampling.

ADASYN [22] is a variation of SMOTE which uses a weighted distribution for different minority class examples according to their difficulty in learning. More synthetic data is generated for minority class examples that are more difficult to learn compared to those that are easier to learn.

2.2.2 *Under-sampling*

Random under-sampling (RUS) is a technique that randomly eliminates examples in the majority class [5]. However, this technique can remove potentially useful information from the training set [5].

The Tomek link (TL) algorithm [58] examines two examples $\mathbf{x}_i$ and $\mathbf{x}_j$ belonging to different classes. A TL occurs if there is not an example $\mathbf{x}_k$ such that $d(\mathbf{x}_i, \mathbf{x}_k) < d(\mathbf{x}_i, \mathbf{x}_j)$ or $d(\mathbf{x}_j, \mathbf{x}_k) < d(\mathbf{x}_j, \mathbf{x}_i)$. If a TL occurs then either one of the examples is noise or it is a borderline case [5]. This information can then be used to under-sample the majority class [5].

One-sided selection (OSS) [34] applies TL to remove borderline and noisy majority class examples. It then applies the condensed nearest neighbour (CNN) rule to remove examples far from the decision border by finding a consistent subset of majority samples. A subset D of E is consistent if a 1-NN classifier trained with D correctly classifies E [5]. The process draws one random majority class example as well as all minority class examples and puts them in D [5]. Every misclassified example is then added from E to D [5].

Neighbourhood cleaning rule (NCL) [36] uses the edited nearest neighbour rule (ENN) to remove majority class examples. ENN removes examples whose label differs from the class of at least two of its nearest 3 neighbours. NCL examines each example $\mathbf{x}_i$ and its 3 nearest neighbours. If $\mathbf{x}_i$ belongs to the majority class and the neighbours contradict this class then $\mathbf{x}_i$ is removed [5]. If $\mathbf{x}_i$ belongs to the minority class and the neighbours contradict this class then the neighbours from the majority class are removed [5].

Instance hardness threshold (IHT) [55] selects the m examples from the majority class that have the highest probability estimates for that class when tested in k-fold cross validation with a classifier (such as SVM with linear kernel). m is the number of samples in the minority class.

NearMiss-1 (NM-1) [64] picks majority class examples which have the smallest average distance to the three nearest minority class examples.

Cluster centroids (CC) applies the k-means algorithm with m clusters to the majority class and uses the coordinates of the cluster centroids as the majority samples. As in ITH, m is the number of samples in the minority class.

2.2.3 *Combination*

SMOTE + TL [4] first applies SMOTE then applies TL. Applying SMOTE can cause samples to extend too deep into the other class's space [4]. TL is used as a data cleaning method to remove examples from both classes to produce well-defined class clusters [4].

SMOTE + ENN works similarly to SMOTE + TL but ENN is more aggressive at removing samples than TL [5].

2.3 Evaluation metrics

We use a number of metrics to evaluate the performance characteristics of the classifiers and balancing algorithms. Each metric measures a different performance criterion.

The number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) are used to define several metrics which are used to compare the results in this paper. The true positive rate (TPR) or sensitivity defines the proportion of actual positives which are predicted as positive. The TPR is calculated as $\frac{TP}{TP+FN}$. The true negative rate (TNR) or specificity defines the proportion of actual negatives which are predicted as negative. The TNR is calculated as $\frac{TN}{TN+FP}$.

Accuracy is typically a poor measure of quality for imbalanced datasets as classifiers tend to bias towards the majority class [5, 12]. Several balanced performance metrics are used instead. Balanced accuracy (BACC) provides an equal weighting in TPR and TNR. It is calculated as $\frac{TPR+TNR}{2}$. Matthews correlation coefficient (MCC) also provides a balanced measure of classification quality and is scored between -1 and 1. The MCC is calculated as: $\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$.

Brier score (BS) measures the accuracy of probabilistic predictions in a range of 0 to 1 where a lower score is better [56]. The BS is calculated as: $\frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$ where N is the number of predictions made, f_t is the probability of the positive class being true and o_t is the actual outcome [56].

The receiver operating curve (ROC) compares TPR for a given FPR. Area under ROC (AUC) is an aggregate measure which averages performance of the classifier over all threshold values. However, recent results suggest that AUC does not treat the relative cost of misclassification the same for each classifier [21]. H-measure is an alternative to AUC that remedies this issue through use of a β-distribution which specifies the relative severity of misclassification in a consistent manner for each classifier [21]. H-measure ranges from 0 for a random classifier to 1 for a perfect classifier [21].

3. RELATED WORK

There have been relatively few studies that apply classification techniques to directly predict defaulters of treatment and CHW attrition. The work that has been done for TB default focuses on determining the individual features associated with treatment default. To get an understanding of the work we want to do, we evaluated the credit scoring field as it is a similar binary classification problem and has been well researched.

3.1 Determining predictors of TB default

There have been many studies which focus on determining the factors associated with TB default. We evaluated a selection of these publications [10, 30, 31, 35, 47, 54]. The majority of techniques use a form of LR to 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 two [54, 35]. Lackey *et al.* [35] only picked individuals who did not have a history of past default. Only one publication [31] found the feature not to be significant to the 95% confidence level. 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.* [54] and Jittimanee *et al.* [31] found distance and time to treatment site to be significant respectively. Other significant features such as illegal drug use [35], use of herbal medication [47], daily jobs [31], history of lung cancer [10] and history of liver disease [10] only appeared once in the datasets. It cannot be discerned if the significance is generalisable or specific to the dataset.

3.2 Credit scoring

The credit scoring field aims to predict high risk individuals who should not be provided credit. This field has been well researched in the last 20 years. We reviewed a selection of these papers to gain insight on classifier performance as well examine their experimental methodology. Lessmanna *et al.* [39] conducted a large scale review of credit scoring papers in the last 10 years as well as their own independent evaluation. Their experimentation found that ensemble techniques performed better than individual classifiers with RF achieving the best results averaged over all the datasets. ANN, LR and SVM (RBF) provided the best results for individual classifiers with ANN beating LR and SVM [39]. This result contradicts several earlier studies which found SVMs to outperform ANNs [14, 26, 29, 41]. Several other papers have also found ensemble techniques to be better than a single classifier [24, 48, 60, 61]. Only one paper reviewed found an ANN performed better than an ensemble of classifiers [59].

Of the credit scoring papers we reviewed [1, 3, 6, 14, 16, 24, 29, 26, 37, 41, 43, 44, 48, 59, 60, 61] all provide measures of the correctness of the classifier such as accuracy, TPR and FPR. However, only a few [16, 29, 44, 61] conduct statistical tests to determine if there is a significant difference between results. Lessmanna *et al.* [39] recommends the Brier test to measure the accuracy of the probability estimates produced by the classifiers. None of the evaluated papers in their review nor our own review used such a test.

4. EXPERIMENTAL DESIGN AND EVALUATION

A systematic experimental design was developed to ensure repeatable results.

4.1 Experimental Approach

We divide our investigation into 5 complementary experiments: first we focus on determining the importance of data balancing and parameter tuning for each classifier; second we transition into benchmarking each classifier and producing recommendations on which are suitable choices; third we focus on determining optimal data balancing schemes; fourth we use feature selection to remove redundant features and determine important features; fifth we determine how the time to default affects patient classification.

All our experiments use stratified $n\times5$-fold validation to divide a single dataset into multiple training and testing datasets. Stratified k-fold divides the dataset into k segments and ensures that each segment contains the same ratio of positive and negative examples as the dataset as a whole. k training and testing sets are created by using one segment as the testing dataset and every other segment as the training data. This is repeated such that every segment is used as testing data. We then repeat the stratified 5-fold validation n times to obtain robustness in the results. 5-fold was used instead of 10-fold to ensure that there was a reasonable sample of positive and negative examples per fold.

To provide fair comparison between results and to facilitate repeatability, all data balancers, stratified k-fold algorithms and classifiers uses set random seeds. The seeds are randomly generated before the start of the execution and differ across each run. This ensures that each result utilises the same training and testing folds per run but that these folds differ across runs. The random seeds are saved in the experiment results to facilitate reproducibility.

All data is first pre-processed before it is used for each experiment. We opted to remove samples with missing data to prevent it causing bias in our classification. Most classification techniques are not equipped to handle categorical data. To address this issue, we use "one-hot encoding" to encode a feature with n categories into n separate binary features corresponding to each category. The feature that corresponds with the categorical value is set to 1 while the other $n-1$ features are set to 0. We standardise all numeric features to ensure that it has zero mean and unit variance since most classifiers expect features to have a normal distribution. For each fold, the mean and standard deviation are derived from the training set and then used to standardise both the training and testing set.

To determine significance between results we follow Demšar's [15] recommendation and use the Friedman test with the Nemenyi post-hoc. We use this approach since classifier results are typically not normally distributed and since we want to determine if there is significance across multiple datasets. The Friedman test has a null-hypothesis that each approach is equivalent to one another and therefore their average ranks are equal [15]. If the null-hypothesis is rejected then a post-hoc test can be conducted to determine if the performance between two individual approaches are equal. The Nemenyi test has been criticised as being overly conservative [19]. However, we prefer this approach as it will be more resilient to minor differences caused by noise and stochastic variance. The statistical tests are given the un-averaged results from each run of the stratified $n \times 5$-fold validation.

4.1.1 Parameter tuning

Each classifier typically has several parameters which can be configured. We conduct parameter tuning to determine its importance for our datasets and its effect on each classifier. We also want to see if just the addition of a data balancing algorithm can improve the classification. We first test each classifier at its default parameters. The ELM does not have default parameters so we set the default ELM to have one hidden layer with 100 neurons using the sigmoid activation function. We chose this as it is very similar to what scikit-learn [49] uses as its default parameters for the ANN. To see the effect of a data balancing algorithm on the classifiers with default parameters, we will use the data balancer that results in the highest BACC. Finally, for each classifier we search a grid of reasonable parameters and select the parameters that yield the highest BACC. The grid includes the parameters of the classifier as well as the different data balancing techniques.

Over-fitting the classifiers is a concern when applying the

grid search. We follow Hsu *et al.*'s [25] advice and use cross-validation, specifically three runs of stratified 5-fold cross-validation, when applying parameter tuning to reduce the risk of over-fitting. We also keep the parameter grid fairly coarse as it reduces the chance that we get parameters that work perfectly for the set of training and testing sets used in the cross-validation but not elsewhere.

4.1.2 Comparison of classification techniques

This experiment focuses on comparing each optimised classifier in detail. The experiment uses 15 runs of stratified 5-fold validation for robustness. We record several metrics to aid our comparison: BACC, MCC, BS, H-measure, TPR, TNR, time to fit each training fold and the standard deviations of each metric to see how the results vary for each fold. Each metric tests a different aspect of the performance of the classifier. We will recommend classifiers that show promising results across multiple metrics.

4.1.3 Comparison of data balancing algorithms

There are a number of approaches to data balancing as outlined in Section 2.2. We compare each data balancing algorithm to determine suitable choices. In addition, we want to determine if certain data balancing schemes are better suited to particular classification techniques. We use our parameter grid results from Section 4.1.1 to obtain the optimal parameters for each classifier on every data balancing algorithm. As before, the highest BACC is used to determine these optimal parameters. To ensure robust results, 10 runs of stratified 5-fold cross validation is used. We focus on BACC and H-measure as our primary metrics of comparison. BACC is used to determine the improvement in the balanced classification while H-measure assesses the balancer's effect on the overall classification at each threshold value. An increase in H-measure implies an improvement across multiple FPR thresholds suggesting that the overall classification is improved.

4.1.4 Comparison of feature selection algorithms

Feature selection can be used to help remove noisy features that do not contribute to the overall classification [20]. It can also speed up training times for large datasets [20]. Typically only classifiers that create a model with some linear combination of features can be easily interpreted.

A number of feature selection strategies were selected for this experiment: ANOVA F-value with χ^2 tests, LR, SVM with linear kernel, Bernoulli NB, DT, RF. The feature selection is calculated on the training examples and not the entire dataset to prevent bias in our results [8].

The ANOVA F-test is used on numeric features and tests the null hypothesis that 2 or more groups have the same population mean. χ^2 test is used on categorical features and tests the null hypothesis that a feature is independent of its class and therefore not relevant to the classification. We chose a p-value of 0.1 to reject each of the null hypotheses. This is to try balance feature removal and retention of good features.

We measure our feature selection techniques in two different ways. In the first approach we only retain the 15 most important features. We want to identify the technique that can retain the most important features for our different classifiers and will therefore result in the lowest reduction in BACC. We test this as the papers identified in Section 3.1 typically make use of LR to identify the predictors of default. Our experiment will determine if this is the best approach. We chose 15 features as we deem it a good balance between achieving a reasonable BACC (at least above 0.5) on all our datasets and minimising the number of features. By using 15 features all our datasets will have a large majority of their features removed.

In the second approach we apply recursive feature elimination (RFE) in order to identify the optimal number of features for the particular technique. RFE removes the n least important feature at each iteration until only k features remain. Good feature selection techniques should be able to reduce the overall features while retaining or even improving the classification. For each feature selection technique the set of features that result in the highest BACC are retained. For some training folds, this could result in no features being removed.

To test the feature selection strategies we record the BACC of each classifier with each feature selection strategy. For the RFE approach the minimum features, maximum features and average features selected is also recorded since each fold of the k-fold cross-validation could have a different number of features selected.

4.1.5 An analysis of time to default in the Lima Peru dataset

Since our TB dataset only contains static features that can be captured at registration, we want to determine how the time for an individual to default affects their classification "profile". We use several default ranges (measured in days): 0-30, 0-60, 0-100, 0-200, 50-100, 100-200, 200-1000 and 300-1000. The non-defaulters are randomly divided into two sets. The first set is joined with the defaulters in the range and used as the training set while the other set is joined with the defaulters outside the range and used as the testing set. This process is repeated over 100 runs and the results are then averaged. If the time to default does not play a large role in the classification then we expect to see similar results for each default range. If it does play a large role then we expect to see poor classification results and variability in results between default ranges.

4.2 Datasets

The experiments were run on a set of real-world datasets: a TB default dataset from Lima Peru [35], a CHW attrition dataset from India and two credit scoring datasets. The TB dataset was obtained from the Dryad digital repository. The CHW attrition dataset was obtained with permission from Dimagi. The two credit scoring datasets (later referred to as the German and Australian dataset) were obtained from the UCI machine learning repository [42]. The datasets are fully anonymised and ethical clearance was obtained from the University of Cape Town for this work. Table 1 provides an overview of the characteristics of the datasets. The Peru TB dataset is highly imbalanced which makes it ideal to test different balancing algorithms. The German dataset and Indian attrition dataset are slightly imbalanced while the Australian dataset is balanced. The Peru data set came with values pre-discretized and therefore only contains categorical features. The Indian attrition dataset contains 90 days of CHW evaluation measures as well as the project and sector which the CHW is part of. The classification label is whether the CHW attrited before the second quarter.

5. SOFTWARE DEVELOPMENT

This section outlines the software development methodology and details of the implementation.

5.1 Development principles and methodology

Throughout this project we strived to ensure that our software was highly configurable and modular so that the core components could be re-used for each experiment. Our goal was to create a system that could support multiple datasets and classifiers in a generic manner. We wanted the system to handle the entire experimental process including preprocessing and result visualisation. We followed an iterative development methodology for our software. We used feedback from our weekly meeting to improve our experimental methodology and visualisation of results.

5.2 Implementation details

We developed the system in Python 2.7 since it has widespread library support and facilitates quick development. All preprocessing operations such as removal of samples with missing values, creation of dummy variables and scaling of data are executed within the code base. Each experiment was also multi-threaded to reduce the execution time.

To support a wide variety of techniques, we used several pre-existing libraries for our classification and data balancing techniques. We used scikit-learn [49] for most of the classification techniques with the exception of ELM and CLC. Akusok *et al.*'s [2] ELM toolbox was used for the implementation of the ELM. The CLC implementation was obtained from the original authors [13] as a compiled C++ executable. By creating a wrapper, we can support classifiers written in different languages or which have a different interface. We used wrappers to provide a scikit-learn interface for the ELM and CLC classifier. In addition, the CLC classifier expects a tab separated values (TSV) file as input and produces the classification model and predictions as a TSV file. The wrapper is used to accommodate this.

Different data balancing techniques can be used per classifier. For the data balancing algorithms, we used Lemaître *et al.*'s [38] imbalanced-learn package to provide a scikit-learn compatible implementation. We used the PMCMR [50] package in R for our statistical tests because of the lack of availability of a suitable package in Python.

6. RESULTS AND DISCUSSION

This section presents and discusses the results of the experiments outlined in Section 4.1.

6.1 Parameter Tuning

The results of the parameter tuning experiment are summarised in Figure 1. The CLC classifier was excluded as the it would crash when applied to the Lima TB dataset without any balancing technique. For all the datasets, there is a net gain in BACC from using a data balancer. However, as expected, the Australian credit dataset which is already balanced, only sees very minor improvements. Gaussian NB is the only classifier which sees a large improvement for the Australia dataset. This is likely because the IHT balancer which was used with it was able to remove samples which are noisy for building the classifier model.

We see the largest improvement in BACC on the Lima TB dataset, incorporating a data balancing technique greatly improves the BACC for all but one classifier. The exception is Bernoulli NB where there is no difference in BACC for each dataset. We can only hypothesise that perhaps there is some internal class balancing occurring automatically given the balanced classification by default. We could not not find any literature on this behaviour either.

Across all of our datasets, we see marginal improvements on average from tuning our classifiers. Typically the improvement is less than 0.05 when using the default parameters with a data balancer. On the Lima TB dataset the ANN sees an improvement of 0.09 with the tuned parameters over using default with a data balancer. The ELM sees large improvements when tuned on Lima TB and Indian attrition dataset. These improvements are expected because the parameters for these classifiers affect the entire architecture of the model. The SVM with RBF kernel sees large gains in the Indian attrition dataset by increasing the C parameter. The SVM with polynomial kernel sees a large improvement on the Indian attrition and credit scoring datasets. This is because scikit-learn sets the gamma parameter as $\frac{1}{\text{features}}$ by default, which is too low for the polynomial kernel.

A classifier that has minor differences between default with balancer and tuned BACC across each dataset is preferred. It shows that the classifier will always produce the best results it can without the user having to manually identify the optimal parameters for each dataset or subset of a dataset that is used to answer a particular research question. LR, AdaBoost, RF, Bernoulli NB and SVM with linear kernel showed similar results between default with balancer and tuned across all our datasets. If these classifiers show promising classification results compared to the other classifiers then they will be recommended.

6.2 Classifier results

Table 2 and 3 contain the results for each classifier on the Lima TB and Indian attrition datasets respectively. Results for the credit datasets can be found in the supplementary material. SVM with linear and polynomial kernel tended to halt occasionally when probabilistic output was enabled for the Indian attrition dataset so we disabled it. Results for BS and H-measure are unavailable for those classifiers in this experiment. The CLC classifier does not support probabilistic output so its results do not contain those metrics either.

For all our datasets, ANN and LR produce high quality classification. LR has two benefits over an ANN, it produces a white-box model which can be interpreted to understand the classification process [17] and its training time is orders of magnitude faster than an ANN. As the ANN can produce a non-linear model it should be able to produce results better than those that are limited to a linear model. However, our statistical tests outlined in Table 4 indicate that LR is not significantly worse in BACC than the ANN across the datasets. It is possible that our datasets are too small and contain too few features to truly utilise the ANN to its full potential.

Bernoulli NB produced impressive results for the Indian attrition dataset. A Friedman test ($\chi^2_{12} = 739, p < .01$) and post-hoc test revealed no significant difference in BACC between the ANN and Bernoulli NB ($p = .97$). Bernoulli NB has the added benefit that it scales well with data size and number of features. It was the fastest on the attirion dataset which was the largest at 3840 entries per fold and 270 features. The classification performance of Bernoulli NB was

Table 1: Data set summary

Data set	Entries	Num of numerical features	Num of categorical features	Num of binary features	Data balance ratio (Negative:Positive)
Lima TB	1186	0	6	8	8.65:1
Indian Attrition	4801	90	2	0	1.9:1
German Credit	1000	7	11	2	2.33:1
Australian Credit	690	6	5	3	1.25:1

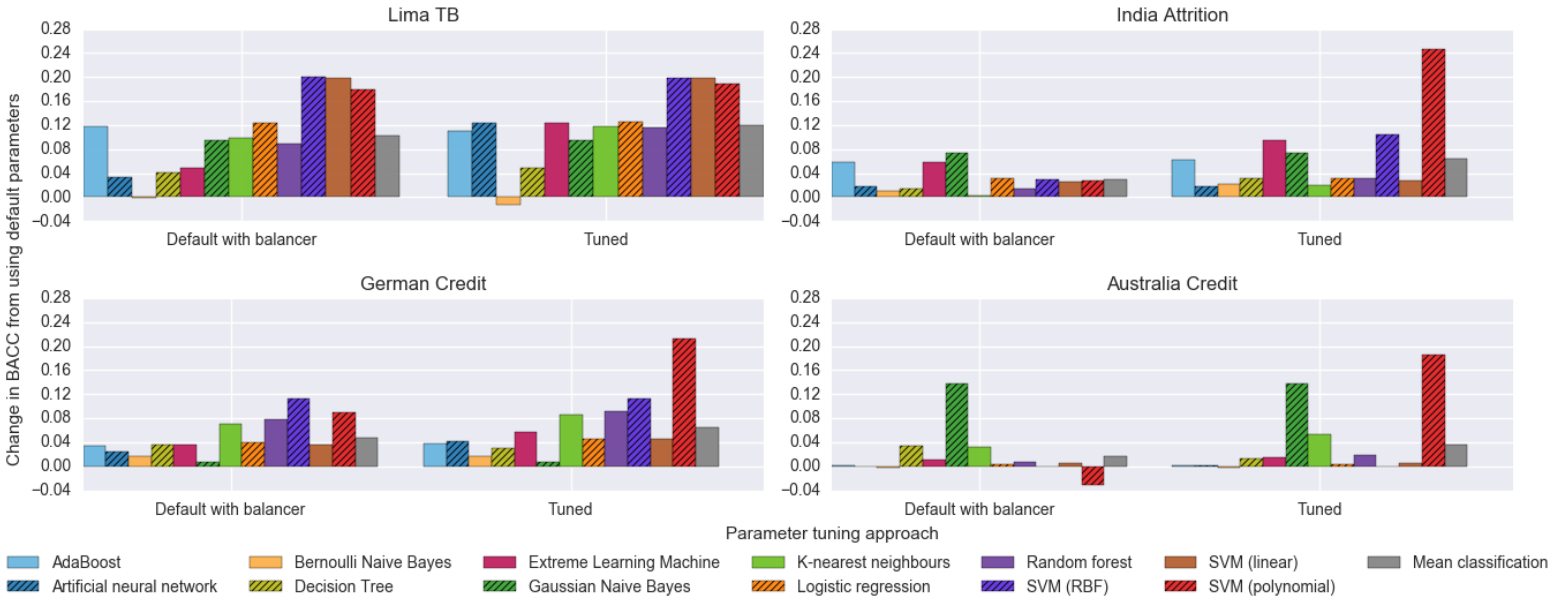

Figure 1: Change in BACC from using default parameters with different parameter tuning approaches. A higher result is better.

Table 2: Lima TB metrics* for each classifier

Classifier	BACC	MCC	BS	H-measure	TPR	TNR	Time to fit† (s)
ANN	**0.7109 (0.0432)**	0.3044 (0.0601)	**0.1740 (0.0414)**	0.3392 (0.0864)	0.6434 (0.0918)	0.7785 (0.0319)	0.3552 (0.0547)
LR	0.7103 (0.0438)	0.3028 (0.0602)	0.1743 (0.0120)	0.3375 (0.0877)	0.6439 (0.0943)	0.7768 (0.0305)	0.0058 (0.0004)
SVM (linear)	0.7006 (0.0419)	0.2969 (0.0569)	n/a	n/a	0.6061 (0.0987)	0.7951 (0.0374)	0.2178 (0.0287)
AdaBoost	0.6998 (0.0453)	0.2756 (0.0604)	0.2411 (0.0019)	**0.3493 (0.0867)**	0.6683 (0.0975)	0.7314 (0.0355)	0.0870 (0.0440)
Bernoulli NB	0.6966 (0.0471)	0.2810 (0.0654)	0.1887 (0.0187)	0.3143 (0.0901)	0.6270 (0.0978)	0.7663 (0.0339)	0.0007 (0.0001)
RF	0.6964 (0.0494)	0.2843 (0.0705)	0.1916 (0.0135)	0.3165 (0.0865)	0.6191 (0.1071)	0.7737 (0.0457)	0.0446 (0.0053)
ELM	0.6953 (0.0515)	0.2811 (0.0714)	0.1836 (0.0140)	0.3130 (0.0843)	0.6199 (0.1100)	0.7707 (0.0401)	0.0050 (0.0010)
SVM (RBF)	0.6952 (0.0441)	0.2821 (0.0612)	0.1944 (0.0139)	0.3232 (0.0848)	0.6195 (0.1060)	0.7709 (0.0496)	0.0114 (0.0007)
SVM (poly)	0.6948 (0.0501)	**0.3054 (0.0738)**	n/a	n/a	0.5608 (0.1092)	**0.8288 (0.0434)**	0.0018 0.0001
Gaussian NB	0.6777 (0.0769)	0.2561 (0.1078)	0.3021 (0.1664)	0.3022 (0.0889)	**0.7002 (0.1422)**	0.6551 (0.2359)	0.0012 (0.0001)
DT	0.6606 (0.0599)	0.2402 (0.0931)	0.2164 (0.0217)	0.2144 (0.1058)	0.5507 (0.1290)	0.7704 (0.0884)	0.0005 (0.0001)
*k*NN	0.6592 (0.0478)	0.2253 (0.0642)	0.2260 (0.0129)	0.2053 (0.0778)	0.5673 (0.1071)	0.7511 (0.0348)	**0.0003 (0.0000)**
CLC	0.5196 (0.0789)	0.0264 (0.1057)	n/a	n/a	0.5428 (0.1592)	0.4964 (0.1503)	0.0695 (0.0060)

*Standard deviation in brackets
†Time to fit individual fold

Table 3: Indian attrition metrics* for each classifier

Classifier	BACC	MCC	BS	H-measure	TPR	TNR	Time to fit† (s)
RF	**0.8276 (0.0117)**	**0.6220 (0.0233)**	**0.1322 (0.0042)**	**0.5619 (0.0257)**	**0.8666 (0.0172)**	0.7886 (0.0187)	1.2816 (0.0138)
Bernoulli NB	0.8231 (0.0138)	0.6200 (0.0266)	0.1330 (0.0089)	0.5355 (0.0285)	0.8314 (0.0248)	0.8147 (0.0180)	**0.0188 (0.0043)**
ANN	0.8182 (0.0138)	0.6123 (0.0280)	0.1363 (0.0101)	0.5261 (0.0285)	0.8208 (0.0278)	0.8156 (0.0261)	3.8548 (1.0586)
LR	0.8051 (0.0125)	0.5788 (0.0238)	0.1364 (0.0055)	0.5063 (0.0238)	0.8399 (0.0233)	0.7704 (0.0182)	0.2717 (0.0114)
AdaBoost	0.8035 (0.0140)	0.5989 (0.0273)	0.1623 (0.0039)	0.4882 (0.0294)	0.7545 (0.0256)	**0.8524 (0.0163)**	6.3176 (0.2101)
Gaussian NB	0.8035 (0.0176)	0.5800 (0.0367)	0.2015 (0.0223)	0.4304 (0.0470)	0.8190 (0.0243)	0.7879 (0.0356)	0.0276 (0.0039)
DT	0.7974 (0.0132)	0.5664 (0.0259)	0.1610 (0.0093)	0.4489 (0.0297)	0.8169 (0.0238)	0.7778 (0.0210)	0.0236 (0.0061)
SVM (linear)	0.7968 (0.0137)	0.5632 (0.0258)	n/a	n/a	0.8275 (0.0270)	0.7661 (0.0191)	7.9526 (1.1028)
SVM (RBF)	0.7853 (0.0148)	0.5456 (0.0288)	0.1556 (0.0078)	0.4540 (0.0304)	0.7896 (0.0243)	0.7811 (0.0195)	27.6928 (0.5112)
*k*NN	0.7822 (0.0137)	0.5546 (0.0268)	0.1455 (0.0073)	0.4571 (0.0280)	0.7327 (0.0266)	0.8317 (0.0179)	0.0661 (0.0055)
SVM (poly)	0.7795 (0.0131)	0.5357 (0.0256)	n/a	n/a	0.7769 (0.0231)	0.7822 (0.0195)	3.2300 (1.1896)
ELM	0.7694 (0.0146)	0.5135 (0.0289)	0.1577 (0.0072)	0.4159 (0.0284)	0.7778 (0.0242)	0.7610 (0.0224)	0.2879 (0.0052)
CLC	0.5082 (0.0115)	0.0362 (0.0449)	n/a	n/a	0.6380 (0.4562)	0.3784 (0.4516)	2.1972 (0.3599)

*Standard deviation in brackets
†Time to fit individual fold

Table 4: Nemenyi post-hoc results† against the best ranked classifier* using all dataset results

Classifier	BACC	MCC	BS	H-measure	Time to fit
ANN	–	–	**1.0**	**0.88858**	< 0.0001
Bernoulli NB	**0.23333**	**0.26715**	< 0.0001	< 0.0001	**0.51226**
RF	**0.99978**	**0.99998**	–	–	< 0.0001
Gaussian NB	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.04533
AdaBoost	< 0.0001	0.00016	< 0.0001	< 0.0001	< 0.0001
LR	**0.71414**	**0.22798**	**0.0819**	< 0.0001	< 0.0001
SVM (linear)	0.00438	0.00748	n/a	n/a	< 0.0001
DT	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.00951
SVM (RBF)	< 0.0001	< 0.0001	0.0033	< 0.0001	< 0.0001
ELM	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
kNN	< 0.0001	< 0.0001	< 0.0001	< 0.0001	–
SVM (poly)	< 0.0001	< 0.0001	n/a	n/a	< 0.0001
CLC	< 0.0001	< 0.0001	n/a	n/a	< 0.0001
Friedman	$\chi^2_{12} = 1274^{\ddagger}$	$\chi^2_{12} = 1179^{\ddagger}$	$\chi^2_9 = 1576^{\ddagger}$	$\chi^2_9 = 951^{\ddagger}$	$\chi^2_{12} = 3117^{\ddagger}$

*Represented with a "–"
†Displayed as p-value with values bolded if they are not significant at $p = .05$
‡Null hypothesis of Friedman test rejected ($p < .0001$)

unexpected as it is designed to operate on binary features. For each dataset we already convert categorical features to multiple binary features as part of the one-hot encoding. However, Bernoulli NB has to apply binarizing on numerical features to transform them into binary features. This process could result in the loss of important information. Since our numerical features are scaled to that of a normal distribution, a binarization threshold of 0 is optimal where input greater than 0 becomes 1 and input lower becomes 0. The classifier also showed the most stable results in our parameter tuning experiment indicating that minimal tuning is required. Bernoulli NB is therefore well suited for rapid experimentation and prototyping of ideas because of its fast training times and solid classification performance.

Huang *et al.* [27, 28] found the ELM produced results similar to that of SVM (RBF) and ANN. For our datasets, the ELM ranked worse than the ANN and SVM (RBF) when measuring with BACC. Our Friedman and post-hoc tests, summarised in Table 4, revealed that the ELM was significantly worse than the ANN ($p < .01$) but was not significantly worse than the SVM (RBF) ($p = .54$). The ELM does deliver on its claim that it can learn much faster than an ANN. It is 13 times faster on the attrition dataset, 64 times faster on the TB dataset, 189 times faster on the German credit dataset and 802 times faster on the Australian credit scoring dataset. However, if learning speed is important then Bernoulli NB may be a better alternative. The Friedman and post-hoc tests summarised in Table 4 found Bernoulli NB to be significantly better in BACC, MCC, H-measure and training speed than the ELM ($p < .01$ for the listed metrics).

We were unable to replicate Luo *et al.*'s [43] results for the CLC on the credit scoring datasets. We were only able to achieve a BACC of 0.6048 and 0.5744 on the German and Australian dataset respectively where Luo *et al.* achieved an accuracy of 0.8480 and 0.8652. Since the Australian dataset is balanced, BACC is the same as accuracy. On the Lima TB dataset and Indian attrition dataset we were unable to achieve results better than random with the CLC classifier. Luo *et al.*'s accuracy of 0.8480 on the German credit scoring dataset is far higher than that seen in the other credit scoring literature utilizing the dataset. The literature typically obtains an accuracy of 0.77 for the dataset [26, 48, 59, 61]. Since the CLC binary provided by the original authors [13] calculated the accuracy of the prediction to terminal; we were able to verify that we were reading the output classification file correctly by calculating the accuracy ourselves and checking that it matched the terminal output. We verified that we had the data formatted correctly by changing the expected output labels on the test dataset to something other than 0 or 1 and observing an accuracy of 0% from the CLC classifier.

Table 4 contains the p-values from the Nemenyi multiple comparison test with the best ranked classifier for each metric. We reject the null-hypothesis at a p-value < 0.05. There is no significant difference between ANN, Bernoulli NB, RF and LR for multiple metrics. Our results are similar to Lessmanna *et al.* [39] who also found ANN, RF and LR to produce good results.

6.3 Data Balancing

Table 5 contains the data balancer results for each classifier on the Lima TB dataset. The Lima TB dataset with an imbalance ratio of 8.65:1 saw great gains in both BACC and H-measure for most of the classifiers. DT, LR, RF and SVM have the ability to conduct internal class balancing via an input parameter. These classifiers saw little variation in both BACC and H-measure across all our datasets. However, the ability to conduct internal class balancing is dependent on the classifier implementation. On the Lima TB dataset, the classifiers which required a balancer to produce competitive results were: AdaBoost, ANN, ELM and kNN. SVM (poly) and SVM (RBF) also saw substantial gains even though they had internal balancing available. The increase in H-measure implies that we are improving the classification across the range of FPR threshold values and not just "moving" to a more balanced FPR threshold value on a ROC curve. On the Lima TB dataset we see greater improvement from over-sampling techniques such as ADASYN and ROS over the under-sampling techniques. Batista *et al.* [5] also observed that datasets with high imbalance ratios performed better with over-sampling techniques. Our results for ROS also align with Batista *et al.* [5] which found it to produce results on par with other over-sampling algorithms. ADASYN produced the highest ranked BACC and H-measure on average. We conducted a Friedman test ($\chi^2_{13} = 760$, $p < .01$) excluding DT, LR, SVM (linear) and RF since they did not improve from the use a balancer. The post-hoc test revealed that all other balancers were significantly worse in BACC than ADASYN ($p < .01$). The Friedman test on H-measure results ($\chi^2_{13} = 428$, $p < .01$) excluded the aforementioned classifiers as well as AdaBoost since not all the base classifiers had support for probabilistic predictions. The post-hoc test also found that all other balancers were significantly worse than ADASYN.

The Indian attrition and German credit datasets each contain a similar ratio of imbalance at 1:1.9 and 1:2.33 respectively. The complete results for both datasets are available in the supplementary material. We excluded the Australian dataset since it is balanced and the only improvement was already highlighted in Section 6.1. The results for H-measure were similar for each data balancer in the Indian attrition dataset (less than 0.01 difference) except SVM (RBF) and SVM (polynomial). This suggests that in reality it is equivalent to shifting to a more balanced FPR threshold but does not improve the overall classification. Both SVM (RBF) and SVM (polynomial) saw similar BACC but improved H-measure (0.02 for both) using SMOTE and SMOTE+ENN respectively. The Indian attrition dataset saw minimal im-

Table 5: Lima TB data balancing results (BACC/H-measure) comparing each classifier to each data balancer

Data Balancer	AdaBoost	ANN	Bernoulli NB	DT	ELM	Gaussian NB	kNN	LR	RF	SVM (linear)	SVM (poly)	SVM (RBF)
None	.597/-	.603/.239	.701/.330	.660/.190	.589/.249	.574/.321	.540/.152	.708/**.341**	.684/.321	.700/.330	.554/.300	.646/.316
CC	.704/-	.698/.329	.695/.314	.631/.163	.681/.272	.547/.274	.677/.267	.690/.316	.589/.215	.689/.308	.683/.309	.685/.248
EEN	.676/-	.660/.329	**.712**/.329	.659/.218	.651/.287	.597/.315	.602/.168	.687/.340	.684/.311	**.709**/.347	.616/.298	.709/**.347**
IHT	.666/.298	.638/.275	.674/.325	.602/.132	.621/.208	.645/.251	.666/.258	.662/.305	.645/.290	.686/.313	.685/.304	.651/.279
NM-1	.666/.289	.653/.278	.650/.289	.615/.105	.628/.165	.538/.130	.649/.223	.659/.286	.624/.182	.661/.290	.669/.290	.669/.287
NCR	.670/-	.648/.327	**.712**/**.333**	.658/.204	.639/.289	.594/.316	.593/.141	.689/.343	**.696**/**.324**	.707/.335	.632/.299	.694/.321
OSS	.622/.171	.603/.212	.707/**.333**	.659/.190	.601/.250	.579/.321	.560/.160	.706/.338	.689/.320	.701/.328	.617/.300	.696/.336
RUS	.701/-	.698/.337	.696/.331	.663/.193	.697/**.328**	.583/.293	.676/.234	.700/.331	.692/.306	.695/.321	.688/.316	.701/.327
TL	.627/.178	.611/.276	.703/.330	.646/.178	.599/.251	.581/.319	.548/.155	.705/.336	.688/.317	.698/.337	.580/.300	.645/.315
ADASYN	.710/-	**.712**/.339	.690/.331	.656/.178	.695/.301	**.671**/.314	.664/.283	**.713**/.338	.690/.301	.704/.332	**.711**/**.349**	**.711**/**.347**
ROS	**.711**/**.344**	.709/.339	.694/.329	.652/.191	**.700**/.323	.547/**.322**	**.682**/**.293**	.704/.333	.688/.306	.701/.322	.623/.265	.691/.315
SMOTE	.701/.319	**.712**/.338	.696/.320	.654/.171	.642/.215	.548/.298	.658/.215	.696/.340	.683/.321	.706/.346	.685/.164	.685/.155
SMOTE+ENN	.705/-	.710/**.342**	.699/.318	.653/.189	.639/.219	.564/.306	.679/.279	.702/.327	.684/.328	.703/**.349**	.521/.315	.690/.314
SMOTE+TL	.701/.319	.705/.336	.696/.320	**.663**/**.211**	.642/.214	.548/.298	.658/.215	.697/.339	.680/.321	.706/.346	.672/.178	.690/.313

provement in BACC. An improvement of 0.03 was obtained with SMOTE for ELM and 0.07 with SMOTE+EEN for Gaussian NB respectively. No other classifier improved by more than 0.01. The German credit dataset sees improvement in H-measure for the ANN and kNN of 0.03 using ROS and 0.07 using IHT respectively. The other classifiers did not see an improvement in H-measure over 0.01. NCR was best for boosting BACC for AdaBoost, ANN, ELM and kNN. Each improved by more than 0.04. No other classifiers were greatly improved. On average, ROS produced the highest ranking BACC and H-measure across the two datasets for each classifier excluding DT, LR, SVM (linear) and RF. A Friedman test ($\chi^2_{13} = 2141$, $p < .01$) and post-hoc test on the BACC results found that only SMOTE ($p = .99$) and SMOTE+TL ($p = .35$) were not significantly worse than ROS. A Friedman test ($\chi^2_{13} = 3023$, $p < .01$) and post-hoc test on the H-measure results found no significant difference between ROS and using no balancer ($p = .99$). In Section 6.1 the balancers had more of an effect as it only used default parameters which were not tuned to improve BACC.

None of our balancer results suggest that certain balancers are always best suited for specific classifiers.

We recommend ADASYN for highly imbalanced datasets especially if one is using a classifier without the ability to conduct internal class balancing such as: AdaBoost, ANN, ELM or kNN. Even classifiers such as SVM (RBF) and SVM (polynomial) with internal balancing benefited from the use of a data balancer. For less imbalanced datasets the improvement is mostly in BACC with little improvement to H-measure. For these datasets the use of a balancer will help balance TPR and TNR but the overall classification at each FPR threshold will likely remain similar. Over-sampling techniques were better overall than under-sampling techniques for the evaluated datasets. We found ADASYN, ROS and SMOTE to be viable choices across the datasets.

6.4 Feature selection

Our feature selection experiment is divided into two sub-goals, identifying the most important predictors and removing unimportant features. Figure 2 (top) contains the results for our first sub-goal. Across our datasets RF and decision tree were able to identify the most important features with the least reduction in BACC. Bernoulli NB is not recommended for selecting the most important features. Our findings suggest that LR may not be the best approach for identifying the most important features.

Our second goal was to use feature selection to remove unnecessary features while maintaining overall BACC. Bernoulli NB only removed a small subset of features from each dataset. SVM (linear) and RF were both able to reduce the number of features substantially while maintaining overall BACC. However, the feature selection was not able to improve BACC in most cases. It is important to determine if this approach reduces the training time significantly, otherwise it serves little purpose. Across our datasets RF reduced training times by 14% and SVM (linear) by 16% on average over no feature selection. We applied the Friedman test ($\chi^2_6 = 5379$, $p < .01$) and post-hoc test which found that all the feature selection approaches resulted in a significant reduction in time to fit ($p < .01$), except the Bernoulli NB approach ($p = .89$), over no feature selection. The difference in training times between the SVM (linear) and RF feature selection approaches was significant ($p = .01$). Overall, we would only recommend applying feature selection for large scale datasets which contain many features where a 10-20% reduction in training time could equate to many hours or days of training time saved. We would recommend SVM (linear) as it provided similar BACC as no feature selection, a large reduction in features and a larger reduction in training time over RF.

6.5 Analysis of time to default on Lima TB dataset

A plot of BACC for each range can be found in the supplementary material. The only default range that showed a large variation in BACC from that of the others was training using the 300+ default range. Using this range resulted in a low TPR when applied to the testing set which contains the other half of non-defaulters and defaulters outside this default range. This suggests that a defaulter's "profile" is similar for those who default within 300 days but different for those who default thereafter. We hypothesise that the incorporation of temporal data, for example from each check-up, would be able to improve the classification for those that default long after registration.

7. LIMITATIONS

While our approach was designed to limit bias and over-fitting as far as possible, the scale of our comparison and size of our datasets could introduce over-fitting nonetheless. For this reason, our results and conclusions incorporate the use of trends, prior knowledge and comparison to our other datasets. We were also limited by the datasets that were available to us for this study. The number of samples in our datasets was small compared to what is available in a production environment which limited our ability to see how the results scale to much larger datasets. Static data, captured at registration, was only available for our TB dataset with no data available from each check-up. Production systems would have this information available which could then be incorporated to improve the classification.

8. CONCLUSIONS AND FUTURE WORK

We performed an analysis of several classifiers, data balancers and feature selection approaches for use in predicting

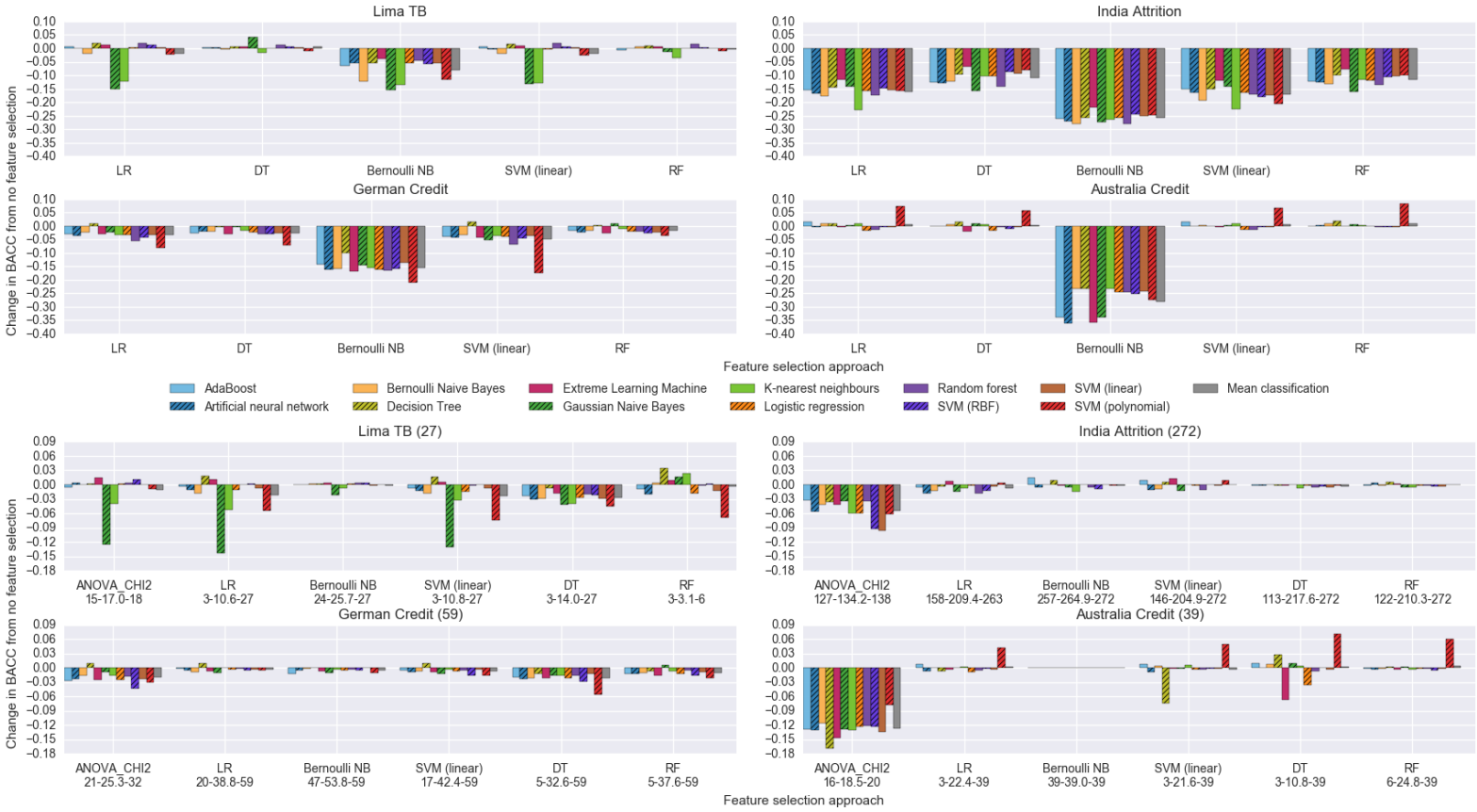

Figure 2: Change in BACC from using no feature selection with different feature selection approaches forced to select 15 features for each classifier (top) and allowed to select any number of features (bottom). A lower result is worse. The Min-Avg-Max features selected is shown in feature selection description with starting number of features in the title (bottom).

treatment defaulters and attrition of CHWs. Since class imbalance is common among the two listed fields, we set out to determine its effect on the classification and tried to address it through the use of data balancing algorithms. We aimed to provide recommendations on the selection of classifier, data balancer and feature selection strategy.

Our experiments found that use of data balancing techniques helped to create a more balanced classification. Parameter tuning only resulted in marginal improvements for most classifiers. The ANN, ELM and SVM (polynomial) saw the greatest improvement in BACC from parameter tuning. We found that the gamma parameter was far too low by default for SVM (polynomial) on 3 of the datasets.

The use of LR, ANN, RF and Bernoulli NB are recommended for several reasons. Bernoulli NB showed above median performance and it can be trained incredibly fast. This allows it to scale well for large datasets and facilitates rapid testing and improvement of one's experimental design. ANN produced excellent results in all our datasets and achieved the highest BACC and MCC across the datasets. LR produced good BACC and MCC in the Lima TB and German credit dataset. It also showed competitive performance in the other two datasets. LR produces a white-box model which allows for easier interpretation of the classification process. RF produced the best BS and H-measure across the datasets. There was no statistical difference between RF and ANN across the datasets in any of the metrics.

The use of a data balancer for AdaBoost, ANN, ELM and kNN was essential for good classification performance on the highly imbalanced Lima TB dataset. Each saw substantial improvements in both BACC and H-measure indicating that the balancer was able to greatly improve the overall classification. We found that over-sampling techniques worked best on the Lima TB dataset with ADASYN being the recommended for highly imbalanced datasets. Classifiers such as LR, RF, DT which can conduct internal class balancing showed strong balanced classification without the need of a balancer. The less imbalanced Indian attrition did not see large improvements to H-measure from using a balancer. This suggests that the balancer was not able to improve the overall classification and was just equivalent to moving to a new FPR threshold on a ROC curve. The German credit dataset only saw improvements in H-measure for the ANN and kNN. Over-sampling techniques such as ADASYN, ROS and SMOTE showed good results across our datasets. Caution should be taken with ROS though, as it can increase the risk of over-fitting [5].

Most of the papers in Section 3.1 used LR to determine the predictors associated with TB default. Our results suggest that RF is better at identifying important predictors. SVM (linear) showed the best performance when RFE was applied. It was able to reduce the number of features and reduce model training time the most. However, we only saw a 16% reduction in training time and therefore only recommend RFE for large datasets with many features so that the reduction equates to many hours or days in training time saved.

Future work would focus on addressing the limitations outlined in Section 7 by using datasets which are orders of magnitude larger and the incorporation of temporal data into the TB default prediction. Classifiers such as recurrent neural networks could be investigated to allow for better incorporation of this temporal data so that a patient's risk profile can be updated at each follow-up.

9. REFERENCES

[1] H. A. Abdou, M. D. D. Tsafack, C. G. Ntim, and R. D. Baker. Predicting creditworthiness in retail banking with limited scoring data. *Knowledge-Based Systems*, 2016.

[2] A. Akusok, K. M. BjÃűrk, Y. Miche, and A. Lendasse. High-performance extreme learning machines: A complete toolbox for big data applications. *IEEE Access*, 3:1011–1025, 2015.

[3] E. Angelini, G. di Tollo, and A. Roli. A neural network approach for credit risk evaluation. *The Quarterly Review of Economics and Finance*, 48(4):733–755, 2008.

[4] G. E. Batista, A. L. Bazzan, and M. C. Monard. Balancing training data for automated annotation of keywords: a case study. In *WOB*, pages 10–18, 2003.

[5] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, 6(1):20–29, June 2004.

[6] H. A. Bekhet and S. F. K. Eletter. Credit risk assessment model for jordanian commercial banks: Neural scoring approach. *Review of Development Finance*, 4(1):20–28, 2014.

[7] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl. Aggregate features and adaboost for music classification. *Machine Learning*, 65(2):473–484, 2006.

[8] M. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. Wright, J. Wilson, F. Agakov, P. Navarro, and C. Haley. Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific reports*, 5:10312, 2015.

[9] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. KÃűnig. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):493–507, 2012.

[10] 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.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, June 2002.

[12] N. V. Chawla, N. Japkowicz, and A. Kotcz. Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, 6(1):1–6, June 2004.

[13] T.-S. Chen, C.-C. Lin, Y.-H. Chiu, H.-L. Lin, and R.-C. Chen. *A New Binary Classifier: Clustering-Launched Classification*, pages 278–283. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[14] P. Danenas and G. Garsva. Selection of support vector machines based classifiers for credit risk domain. *Expert Systems with Applications*, 42(6):3194–3204, 2015.

[15] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of machine learning research*, 7:1–30, Dec. 2006.

[16] V. S. Desai, J. N. Crook, and G. A. O. Jr. A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, 95(1):24–37, 1996.

[17] S. Dreiseitl and L. Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, 35(5-6):352 – 359, 2002.

[18] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139, 1997.

[19] S. Garcia and F. Herrera. An extension on "statistical comparisons of classifiers over multiple data set" for all pairwise comparisons. *Journal of Machine Learning Research*, 9:2677–2694, 2008.

[20] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3:1157–1182, Mar. 2003.

[21] D. J. Hand. Measuring classifier performance: A coherent alternative to the area under the roc curve. *Machine learning*, 77(1):103–123, Oct. 2009.

[22] H. He, Y. Bai, E. A. Garcia, and S. Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328, June 2008.

[23] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

[24] N.-C. Hsieh and L.-P. Hung. A data driven ensemble classifier for credit scoring analysis. *Expert Systems with Applications*, 37(1):534–545, 2010.

[25] C.-W. Hsu, C.-C. Chang, C.-J. Lin, et al. A practical guide to support vector classification. *Department of Computer Science, National Taiwan University*, 2016.

[26] C.-L. Huang, M.-C. Chen, and C.-J. Wang. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, 33(4):847–856, 2007.

[27] G. B. Huang, H. Zhou, X. Ding, and R. Zhang. Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(2):513–529, April 2012.

[28] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3):489 – 501, 2006. Neural Networks Selected Papers from the 7th Brazilian Symposium on Neural Networks (SBRN '04)7th Brazilian Symposium on Neural Networks.

[29] Z. Huang, H. Chen, C.-J. Hsu, W.-H. Chen, and S. Wu. Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, 37(4):543–558, 2004. Data mining for financial decision making.

[30] 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.

[31] 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.

[32] G. H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 338–345, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[33] J. M. Keller, M. R. Gray, and J. A. Givens. A fuzzy k-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-15(4):580–585, July 1985.

[34] M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *In Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann, 1997.

[35] 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.

[36] J. Laurikkala. Improving identification of difficult small classes by balancing class distribution. In *Proceedings of the 8th Conference on AI in Medicine in Europe: Artificial Intelligence Medicine*, AIME '01, pages 63–66, London, UK, UK, 2001. Springer-Verlag.

[37] T.-S. Lee, C.-C. Chiu, C.-J. Lu, and I.-F. Chen. Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*, 23(3):245–254, 2002.

[38] G. Lemaître, F. Nogueira, and C. K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *CoRR*, abs/1609.06570, 2016.

[39] S. Lessmanna, H. Seowb, B. Baesenscd, and L. C. Thomasd. Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update. In *Credit Scoring and Credit Control XIII*, Edinburgh, UK, 2013.

[40] D. D. Lewis. *Naive (Bayes) at forty: The independence assumption in information retrieval*, pages 4–15. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

[41] S.-T. Li, W. Shiue, and M.-H. Huang. The evaluation of consumer loans using support vector machines. *Expert Systems with Applications*, 30(4):772–782, 2006.

[42] M. Lichman. UCI machine learning repository, 2013.

[43] S.-T. Luo, B.-W. Cheng, and C.-H. Hsieh. Prediction model building with clustering-launched classification and support vector machines in credit scoring. *Expert Systems with Applications*, 36(4):7562–7566, 2009.

[44] R. Malhotra and D. Malhotra. Evaluating consumer loans using neural networks. *Omega*, 31(2):83–96, 2003.

[45] A. McCallum, K. Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48, 1998.

[46] C. Mood. Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, 26(1):67–82, 2010.

[47] 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.

[48] L. Nanni and A. Lumini. An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, 36(2, Part 2):3028–3033, 2009.

[49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[50] T. Pohlert. *The Pairwise Multiple Comparison of Mean Ranks Package (PMCMR)*, 2014. R package.

[51] I. Rish. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. IBM New York, 2001.

[52] R. Rojas. Adaboost and the super bowl of classifiers a tutorial introduction to adaptive boosting, Computer Science Department, Freie Universität, Berlin. 2009.

[53] S. Sandro. Understanding logistic regression analysis. *biochem*, 24(1):12–18, 2014.

[54] 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.

[55] M. R. Smith, T. Martinez, and C. Giraud-Carrier. An instance level analysis of data complexity. *Machine learning*, 95(2):225–256, May 2014.

[56] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128–138, 2010.

[57] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston. Random forest: A classification and regression tool for compound classification and qsar modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958, 2003. PMID: 14632445.

[58] I. Tomek. Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(11):769–772, Nov 1976.

[59] C.-F. Tsai and J.-W. Wu. Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, 34(4):2639–2649, 2008.

[60] B. Twala. Multiple classifier application to credit risk assessment. *Expert Systems with Applications*, 37(4):3326–3336, 2010.

[61] G. Wang, J. Hao, J. Ma, and H. Jiang. A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, 38(1):223–230, 2011.

[62] S.-C. Wang. *Interdisciplinary Computing in Java Programming*, chapter Artificial Neural Network, pages 81–100. Springer US, Boston, MA, 2003.

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

[64] I. Zhang and I. Mani. knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, 2003.