Empirical Study of Bagging Predictors on Medical Data

Guohua Liang and Chengqi Zhang

The Centre for Quantum Computation & Intelligent Systems, FEIT, University of Technology, Sydney NSW 2007 Australia

gliang@it.uts.edu.au,chengqi@it.uts.edu.au

Abstract

This study investigates the performance of bagging in terms of learning from imbalanced medical data. It is important for data miners to achieve highly accurate prediction models, and this is especially true for imbalanced medical applications. In these situations, practitioners are more interested in the minority class than the majority class; however, it is hard for a traditional supervised learning algorithm to achieve a highly accurate prediction on the minority class, even though it might achieve better results according to the most commonly used evaluation metric, *Accuracy*. Bagging is a simple yet effective ensemble method which has been applied to many real-world applications. However, some questions have not been well answered, e.g., whether bagging outperforms single learners on medical data-sets; which learners are the best predictors for each medical data-set; and what is the best predictive performance achievable for each medical data-set when we apply sampling techniques. We perform an extensive empirical study on the performance of 12 learning algorithms on 8 medical data-sets based on four performance measures: True Positive Rate (*TPR*), True Negative Rate (*TNR*), Geometric Mean (*G-mean*) of the accuracy rate of the majority class and the minority class, and *Accuracy* as evaluation metrics. In addition, the statistical analyses performed instil confidence in the validity of the conclusions of this research.

Keywords: imbalanced class distribution, medical data, bagging predictors and binary classification.

1 Introduction

Finding effective learning methods and improving prediction accuracy are essential goals for most machine learning approaches (Quinlan 1996), and this is especially true for real-world medical applications. Bagging (Breiman 1996) is a simple and effective ensemble learning method. Due to its promising capabilities in improving accuracy of classification prediction over unstable single learners (Breiman 1996), it has been widely used in many applications. The effectiveness of bagging has been investigated empirically and it has been demonstrated that bagging is very effective for decision trees (Quinlan 1996, Breiman 1996, Bauer and Kohavi 1999, Dietterich 2000, Opitz and Maclin 1999), and Neural Networks (West et al. 2005, Opitz and Maclin 1999, Kim and Kang 2010). Even though the existing studies demonstrate the effectiveness of the bagging predictor, it is not clear whether bagging is superior to single learners in the context of imbalanced medical data-sets, nor which predictor is the best performing learning method on each imbalanced medical data-set.

Our previous works investigate the effectiveness of the bagging predictors in general terms (Liang et al. 2011a) and in imbalanced class distribution terms (Liang et al. 2011b, Liang and Zhang 2011). However, the previous conclusions are based on statistical tests that aggregate the data-sets and do not show which learners are the best prediction models for individual medical data-sets, as various prediction models might behave differently for different kinds of data-sets. They also do not show the best achievable predictive performance for each medical data-set using sampling technique.

In the literature, an empirical study of combined classifiers on medical data (Lopes et al. 2008) compared the performance of three classification methods, C4.5 (Quinlan 1986), bagging, and boosting on 16 medical data-sets and 16 generic data-sets. The evaluation was based on the accuracy of these learning methods as a performance measure; their research did not address the challenging issues of medical data-sets: imbalanced class distribution and the unequal costs of mis-classification errors in different classes. Moreover, accuracy is an inappropriate performance measure for evaluating imbalanced data-sets (Maimon et al. 2010, Chawla et al. 2002).

The majority of medical applications involve learning from imbalanced binary classification data-sets in which the proportion of the class distribution is skewed, the number of instances of the majority class is higher than those of the minority class, and practitioners are more interested in the minority class than the majority class, such as breast cancer early detection, in which the minority class is quite small with an unequal high cost associated with mis-classification errors in different classes. If a patient with breast cancer is mis-classified as normal, the patient will miss the opportunity for his/her earlier stage cancer detection and treatment; while if a patient without breast cancer is mis-classified as having cancer, it will cause unnecessary stress and treatment. Traditional supervised learning algorithms perform poorly in predictive accuracy over the minority class, even though they may produce high overall accuracy (Phua et al. 2004, Ng and Dash 2006, Maloof 2003, Su and Hsiao 2007, Chawla 2010). We therefore employ four measures, True Positive Rate (*TPR*), True Negative Rate (*TNR*), geometric mean (*G-mean*) of the accuracy rate of the majority class and minority class, and *Accuracy* as evaluation metrics to assess the effectiveness of bagging in terms of learning from medical data-sets.

To solve the problem of imbalanced class distribution and increase the *Accuracy* of the prediction model, the most commonly used methods are sampling-oriented methods and algorithms-oriented methods (Liu and Chawla 2011).

In this study, we utilize under-sampling techniques to investigate the performance of bagging predictors at different levels of class distribution and report the best achieved performance of bagging by using sampling techniques based on the *G-mean* evaluation metrics.

The main objectives of this paper are threefold: we (1) determine whether bagging is superior to single learners in the context of imbalanced medical data-sets, (2) determine which learners give the best performance on each medical data-set with natural class distribution, and (3) report the best achieved performance of the bagging predictors on each medical data-set by using sampling techniques.

The paper is organized as follows. Section 2 presents details of the designed framework. Section 3 presents sampling techniques and Section 4 presents the evaluation metrics. Section 5 presents the experimental setting and Section 6 presents the experimental results analysis. Section 7 concludes the paper.

2 Designed Framework

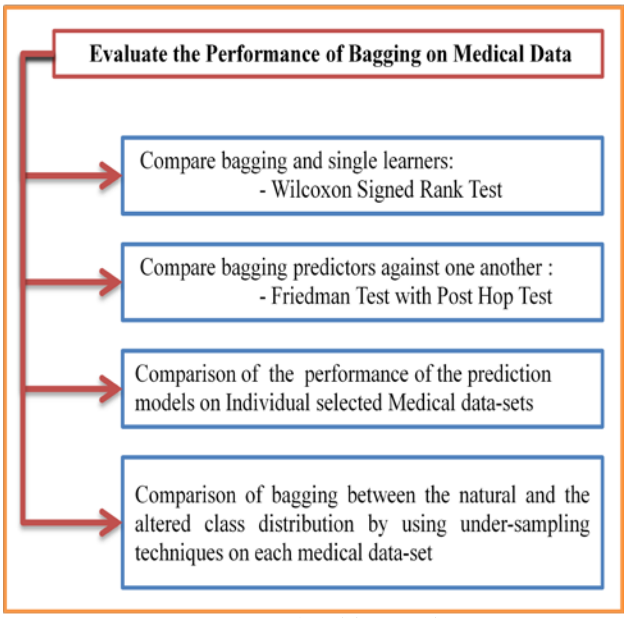

Evaluate the Performance of Bagging on Medical Data
Compare bagging and single learners:
- Wilcoxon Signed Rank Test
Compare bagging predictors against one another :
- Friedman Test with Post Hop Test
Comparison of the performance of the prediction models on Individual selected Medical data-sets
Comparison of bagging between the natural and the altered class distribution by using under-sampling techniques on each medical data-set

Figure 1: Designed framework

The designed framework and the evaluation of bagging predictors on Medical Data-Sets are broken down into four tasks as follows:

- Compare bagging predictors with single learners: the Wilcoxon Signed Ranks Test is used to compare two learners to determine whether bagging outperforms a single learner on medical data-sets.
- Compare the performance of bagging predictors against one another: the Friedman test with the corresponding Post-hoc Nemenyi test is used to compare multiple learners to determine which bagging predictors have the best performance over all 8 imbalanced data-sets,
- Compare the performance of the prediction and report the best performance models with the natural class distribution on each individual medical data-set based on four evaluation metrics: *G-mean, TPR, TNR* and *accuracy rate*.
- Compare the performance of bagging predictors between the natural class distribution and the altered levels of class distribution to determine the best performance of the bagging predictor on each medical data-set.

3 Sampling Techniques

Sampling techniques are commonly used to improve the performance of the prediction model for imbalanced data-sets (Chawla et al. 2002, Chawla et al. 2003, Weiss and Provost 2003), e.g., under-sampling and over-sampling SMOTE (Chawla et al. 2002), Borderline-SMOTE (Han et al. 2005), and Safe-Level-SMOTE (Bunkhumpornpat et al. 2009).

We utilize under-sampling techniques to vary the class distribution of the data to investigate the performance of bagging predictors over medical data-sets, i.e., to alter each original imbalanced data-set, D with sample size M into nine new data-sets, D_1, D_2 ... D_9 with new sample size M_1, M_2 ... M_9, respectively.

We consider the entire minority class samples as a positive class (P) and the proportions of P are as follows: $P = 10\% M_1 = 20\% M_2 = ... = 90\% M_9$, respectively. Then we select the majority class randomly without replacement as a negative class (sample size N_1, N_2 ... N_9), and the proportions of the negative class are as follows: $N_1 = 90\% M_1$; $N2 = 80\% M_2$ ... $N_9 = 10\% M_9$, respectively to form the new data-sets, D_1, D_2 ... D_9. Each original imbalanced data-set D is thereby altered into nine different levels of class distributions.

10 trials 10-fold cross-validation is performed on each of the new data-sets, D_1, D_2 ... D_9, so that the test-set has the same distributions as the training-set. We then compare the results of *G-mean* from nine different class distributions on each medical data-set and report the best results achieved on each data-set using sampling techniques.

4 Evaluation Metrics

Accuracy is a popular choice for evaluating the performance of a classifier; however, it might not be a good metric for measuring the performance of medical data-sets. The challenge issues of the most medical applications are imbalanced class distribution problem and unequal costs of the mis-classification errors in different classes. The minority class is more important than the majority class; normally a high prediction accuracy is required in a minority class and therefore a simple estimated accuracy has limitations in evaluating the performance of a classifier on a minority class (Fawcett 2006). We therefore adopt four measures, *Accuracy*, True Positive Rate (*TPR*), True Negative Rate (*TNR*), and *G-mean* as evaluation metrics.

In this paper, we consider the minority class as the positive class and the majority class as the negative class. Following this convention, *TP* refers to the number of positive instances correctly classified as the positive class; *TN* refers to the number of negative instances correctly classified as the negative class; *FP* refers to the number of negative instances incorrectly classified as the positive class; and *FN* refers to the number of positive instances incorrectly classified as the negative class (Chawla 2010, Guo et al. 2008).

Accuracy (*Acc*) is commonly used as a performance measure of a classifier for balanced learning. However, it has been considered an improper performance measure for evaluating learning from imbalanced data (He and Garcia 2009, Provost et al. 1998, Maloof 2003, Weiss and Provost 2003).

TPR and *TNR* evaluate the performance of a binary classification algorithm directly on the minority class and the majority class respectively. *TPR* refers to the proportion of the minority class that has been correctly classified as a positive class, while *TNR* refers to the proportion of the majority class that has been correctly classified as a negative class. The *G-mean* of the accuracy rate of the majority class and minority class was suggested as a performance measure to assess the effectiveness of learning methods for imbalanced learning (Ng and Dash 2006, He and Garcia 2009, Provost and Fawcett 2001). Table 1 presents the confusion matrix for a binary classification problem. Table 2 presents the formulas of both True Positive Rate and True Negative Rate in the first row, the formula of *G-mean* in the second row, and the formula of Accuracy *(Acc)* in the last row.

Table 1: Confusion matrix for a binary classification problem

	Predicted Positives	**Predicted Negatives**
Positive Instances *(P)*	*True Positive (TP)*	*False Negatives (FN)*
Negative Instances *(N)*	*False Positive (FP)*	*True Negatives (TN)*

Table 2: True Positive Rate, True Negative Rate and G-mean

$TP_{rate} = \frac{TP}{TP + FN}$	$TN_{rate} = \frac{TN}{TN + FP}$
$G - mean = (TP_{rate} * TN_{rate})^{1/2}$	
$Acc = \frac{TP+TN}{TP+TN+FP+FN}$	

5 Brief Overview of Single Learner and Bagging

In this section, we briefly introduce two basic concepts: what constitutes a single learner of supervised learning and what is bagging.

Single learner refers to supervised learning using the labelled samples to form a classifier (called a single learner or prediction model) and having a function that can be used to predict new samples with pre-defined class labels. Figure 2 presents a prediction model of a single learner in supervised learning.

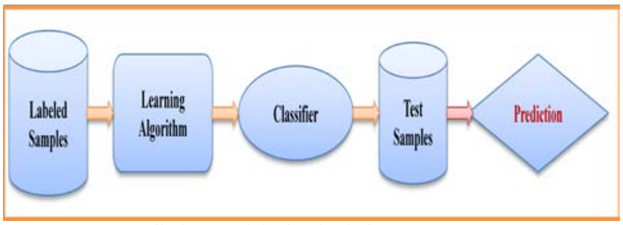


Figure 2: Prediction model of a single learner

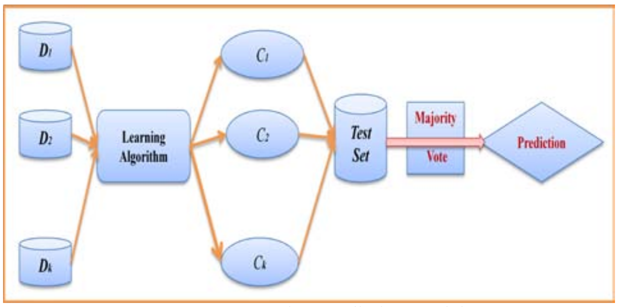
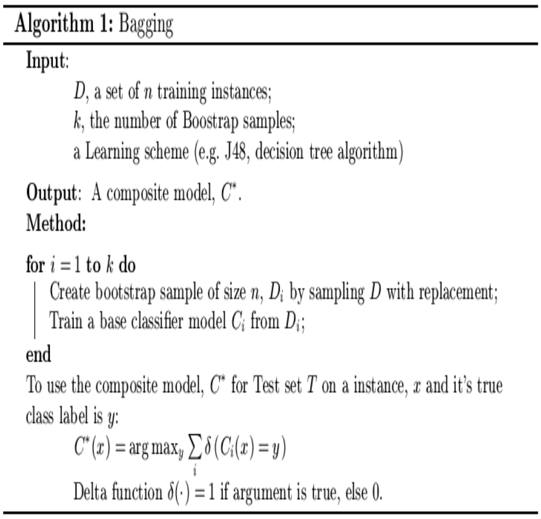


Figure 3: Bagging prediction model

Bagging represents a set of classifiers ($C_{1,}$ *C2*... C_k) (called base learners) which are generated from a set of bootstrap samples ($D_{1,}$ D_2 $_{...}D_k$) to form an ensemble method for prediction, and its function is to predict new samples by a set of classifiers; a final prediction is made by taking a majority vote .

Figure 3 illustrates the basic framework of a bagging prediction model by using bootstrap sampling and voting techniques to improve the performance of the bagging prediction model. Bagging is known as “bootstraps aggregating”. Firstly, for each of the bootstrap samples ($D_{1,}$ D_2 $_{...}D_k$), a new training set D_k is randomly drawn from the original training set *D* of *m* instances with replacement conducted by repeated drawing *m* times. Each bootstrap sample therefore contains the same number of *m* instances as the original training set *D*; some instances may appear many times, while some instances may not appear. Secondly, the *k* bootstrap samples of a training set with *m* instances will generate *k* classifiers ($C_{1,}$ C_2 ... C_k). Finally, the unseen instance *x* of the test set will be predicted by applying each of the *k* classifiers C_i (*i* =1 to *k*) and a final decision *C** is made by majority vote of all classifiers (C_1 ... C_k). The algorithm for bagging is given in Figure 4.



```
Algorithm 1: Bagging
Input:
    $D$, a set of $n$ training instances;
    $k$, the number of Boostrap samples;
    a Learning scheme (e.g. J48, decision tree algorithm)
Output: A composite model, $C^*$.
Method:
for $i = 1$ to $k$ do
    Create bootstrap sample of size $n$, $D_i$ by sampling $D$ with replacement;
    Train a base classifier model $C_i$ from $D_i$;
end
To use the composite model, $C^*$ for Test set $T$ on a instance, $x$ and it's true
class label is $y$:
    $C^*(x) = \arg\max_y \sum_i \delta(C_i(x) = y)$
    Delta function $\delta(\cdot) = 1$ if argument is true, else 0.
```

Figure 4: Algorithm of Bagging (Breiman 1996)

6 Experimental Setting

This section includes three subsections as follows: A. software and parameter settings, B. selection of base learners, and C. data-set selection.

6.1 Software and Parameter Settings

We performed 10-trial 10-fold cross-validations to evaluate bagging and single learners on 8 medical data-sets, which were collected from the UCI Machine Learning Repository (Merz and Murphy 2006). We used WEKA implementation of the 12 algorithms with their default parameter settings in this empirical study (Witten and Frank 2005). We implemented the bagging predictor in Java platform. In order to reduce uncertainty and obtain reliable experimental results, all the evaluations of bagging performance are assessed under the same test conditions by using the same randomly selected bootstrap samples with replacements in each fold of 10-trial 10-folds cross-validation on each data-set.

6.2 Selection of Base Learners

Twelve learning algorithms have been selected for this study. We first select the most commonly used learning algorithms in real-world applications: Support Vector Machines (SVM), Neural Network learner – Multi Layer Proceptron (MLP), Naïve Bayes learner (NB), and K-nearest-neighbours (KNN). We then select rule learners: PART, Decision Table (DTable), and OneR. We finally select tree family learners, C4.5 Decision Tree (J48), DecisionStump (DStump), RandomTree (RandTree), REPTree and Naïve-Bayes-Trees (NBTree).

6.3 Selection of Data-Sets

Table 3: Imbalanced Medical Data-Sets

		Information Data		Class Data		
ID	Name	attribut	instance	frequency	P%	clas
1	breastc	10	286	201,85	29%	2
2	diabetes	9	768	500,268	34%	2
3	heart-c	14	303	165,138	45%	2
4	sick	30	3772	3541,231	6%	2
5	heart-h	14	294	188,106	36%	2
6	stalogHe	14	270	120,150	44%	2
7	wbreastc	10	699	458,241	34%	2
8	WDBC	31	569	212,357	37%	2

A summary of the characteristics of the eight imbalanced medical data-sets is displayed in Table 3. The selected medical data-sets are binary classes. The selection of the eight data-sets covers the number of instances, which varies from small to large up to 3772, the number of attributes, which varies from 9 to 31, and the natural class distribution (*P%*), which indicates the percentage of the positive instances from the total instances of each data-set. The results vary from 6.1%, the extremely imbalanced data-set 'sick' to 45% the almost balanced data-sets 'heart-c' and 'stalogHeart'.

7 Experimental Results Analysis

This section presents the experimental results analysis including four sub-sections as follows: A. comparison of bagging with single learners, B. comparison of bagging predictors on medical data-sets, C. comparison of the performance of 24 prediction models and report the best prediction model on each individual data-set, and D. comparison of the performance of bagging predictors between natural class distribution and the altered class distribution by using under-sampling techniques on each medical data-set.

7.1 Comparison of Bagging and Single learners

This subsection compares bagging and single learners over multiple medical data-sets to determine whether bagging is superior to single learners based on two evaluation metrics, *Accuracy* and *G-mean*.

The Wilcoxon Signed Rank Test is used to compare two learners - bagging and a single learner over multiple data-sets - to determine whether bagging is superior to a single learner.
The Null Hypothesis is that the median of differences between bagging and a single learner equals 0.
Rule: Reject the Null Hypothesis if the p-value Test Statistic W is less than .05 at the 95% confidence level of significance.

Table 4: Compare bagging with each single learner based on Wilcoxon Signed Rank Test on *Accuracy*. The significance level is .05.

Wilcoxon Signed Rank Test on Accuracy						
Learners	J48	RepTree	Randtree	NB	SVM	Dstump
p-value	.025	.012	.012	**.207**	**.138**	**.128**
Learners	OneR	Dtable	PART	KNN	NBTree	MLP
p-value	.012	**.208**	.012	**0.092**	.017	.012

Tables 4 and 5 present the summarized results of the Wilcoxon Signed Rank Test on the evaluation metrics, *Accuracy* and *G-mean* for the comparison of the two learners: single learners versus their corresponding bagging predictors, i.e., we compare bagging J48 and single learner J48. If the p-value is greater than α value, .05, we accept the Null Hypothesis and the p-values are highlighted.

Table 4 indicates that bagging does not perform statistically significantly better than the single learners NB, SVM, Dstump, Dtable and KNN on eight Medical data-sets based on the evaluation metric, *Accuracy*. Table 5 indicates that bagging is statistically superior to the single learners J48, RandTree, OneR, PART and MLP on eight medical data-sets based on the *G-mean* evaluation metric.

Table 5: Compare bagging with each single learner based on Wilcoxon Signed Rank on *G-mean*. The significance level is .05.

Wilcoxon Signed Rank Test on Gmean.						
Learnr	J48	RepTree	Randtree	NB	SVM	Dstump
p-value	.036	**.161**	.036	**.069**	**.093**	**.866**
Learnr	OneR	Dtable	PART	KNN	NBTree	MLP
p-value	.017	**.779**	.036	**.327**	**.484**	.012

7.2 Comparison of the Performance of Bagging Predictors on Imbalanced Medical Data-Sets

Friedman Test and Post-hoc Nemenyi Test: Both tests are non-parametric for comparing multiple algorithms over multiple datasets. Firstly, all the algorithms are ranked on each data-set, giving the best performing algorithm the rank of 1, the second best rank 2, and so on. If there are ties, average values are assigned. Secondly, the average rank of the algorithm is calculated. Finally, the Friedman test compares the average ranks of algorithms and checks whether there is a significant difference between the mean ranks.

The Null Hypothesis of this test states that the performances of all algorithms are equivalent. If the Null Hypothesis is rejected, it does not determine which particular algorithms differ from one another. Because the test result does not show exactly where that significant difference occurs, a post-hoc Nemenyi test is needed for additional exploration of the differences between mean ranks to provide specific information on which mean ranks are significantly different from on another. The critical difference is calculated as:

$$CD = q_{\acute{a}}\sqrt{\frac{d(d+1)}{6N}}$$

Where d is the number of algorithms, N is the number of data-sets, and the critical values q_{α} are based on the Studentized range statistic divided by $\sqrt{2}$. If the mean ranks are different by at least the critical difference, the performance of learners is significantly different. Demšar has presented how to calculate the critical difference of the Nemenyi test in more detail (Demšar 2006).

Table 6: Ranking order of the performance of bagging based on *G-mean* and their Mean Ranks.

Gmean	MLP	NB	NBTree	SVM	PART	RdTree
breastc	2	1	7	6	5	8
diabetes	1	3	2	8	7	5
sick	10	9	5	12	3	7
heart-c	3	2	4	1	5	7
staHeart	3	1	4	2	5	7
heart-h	3	1	4	2	5	6
wdbc	2	10	4	1	3	6
wbreastc	3	1	2	4	6	5
Mean Rank	3.375	3.5	4	4.5	4.875	6.375

	J48	RepTree	Dstump	KNN	Dtable	OneT
breastc	9	11	3	4	12	10
diabetes	4	6	10	11	9	12
sick	1	4	2	11	8	6
heart-c	6	8	10	11	9	12
staHeart	10	6	12	9	8	11
heart-h	10	11	7	8	12	9
wdbc	7	8	11	5	9	12
wbreastc	7	9	12	8	10	11
Mean Rank	6.75	7.875	8.375	8.375	9.625	10.375

Table 6 presents the ranking order of the performance of bagging predictors on each imbalanced medical data-set based on the evaluation metric *G-mean*. Firstly, we divide Table 6 into two parts. In each part, the first row presents the ascending order of the bagging predictors according to their mean rank of the *G-mean* measure in the 10th row. Secondly, the second to ninth rows present the ranking order of the bagging predictors on each individual medical data-set, e.g., bagging MLP performs best on the diabetes data-set ranking as 1, followed by bagging NBTree ranking as 2, and bagging OneR ranked 12 is the worst bagging predictor on the same data-set. The last rows present the mean ranks of the performance of the bagging predictor over all eight medical data-sets. On the other hand, we observe that different bagging predictors behave differently for different medical data-sets, e.g., bagging MLP performs well on most of these medical data-sets, except for sick data-set which is an extremely imbalanced and high dimensional large data-set; bagging NB performs best (ranking as 1) on *four* medical data-sets, breastc, StatlogHeart, heart-h and wbreastc, but performs poorly on the other two data-sets, sick and WDBC, which are high dimensional attributes or extremely imbalanced class distribution data-sets; while bagging J48 and DStump perform well on the extremely imbalanced and high dimensional largest medical data-set, sick.

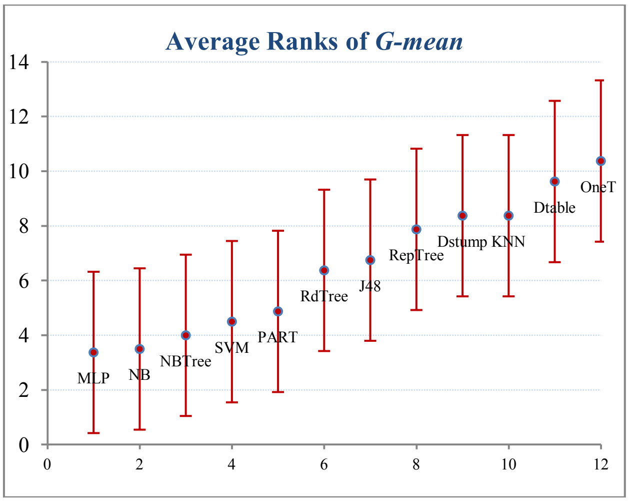


Figure 5: Comparison of all bagging predictors from the Friedman and Post-hoc Nemenyi test, where the x-axes indicate the mean rank of each bagging predictor, the y axes indicate the ascending ranking order of the Bagging predictors, and the horizontal error bars indicate the “critical difference”.

Figure 5 presents the results of the mean ranking of the performance of bagging predictors over all eight medical data-sets based on the Friedman and Post-hoc Nemenyi tests. The results indicate that the group of bagging MLP and NB are the best bagging predictors, while bagging OneR is the worst bagging predictor. The performances of two bagging predictors are significantly different if the horizontal bars do not overlap; therefore, there is a statistically significant difference between the group of two best bagging predictors, MLP and NB and the worst bagging predictor OneR. However, there is not a statistically significant difference between remaining bagging predictors.

7.3 Comparison of the Performance of the Prediction Models on Individual Medical Data-Sets

In this subsection, we compare the performance of the prediction models, bagging predictors and single learners on eight selected medical data-sets. For the data-set selection, we first select breastc data-set which has 10 attributes and 286 instances, in which the proportion of the minority class is 29%; secondly, we select three moderately imbalanced data-sets, WDBC, heart-h, diabetes and wbreastc in which the proportions of the minority class are 37%, 36%, 34% and 34%, respectively; thirdly, we select an extremely imbalanced data-set, sick, which has 30 attributes and 3772 instances, in which the proportion of the minority class is 6%. Finally, we select two almost balanced data-sets, heart-c and stalogHeart data-sets, in which the proportions of the minority class are about 45%.

Figures 5 to 13 inclusive present a comparison of the performance of all the prediction models on eight medical data-sets, breastc, diabetes, sick, heart-h, WDBC, heart-c, wbreastc and statlogHeart. Each graph presents the summarization of the observed performance of the prediction models based on four measures, *G-mean, TPR, TNR* and *Accuracy* on each of the selected data-sets. For each plot, the horizontal axis indicates the ranking order of all the prediction models based on the descending order of the performance measure, *G-mean*, while the vertical axis indicates the value of the four performance measures.

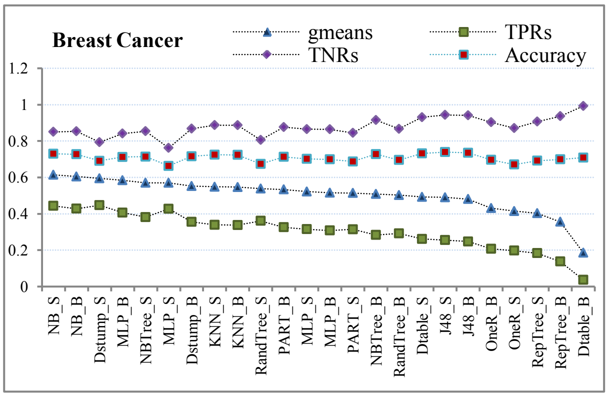


Figure 6: The performance of prediction models on breastc data-set.

Figure 6 shows that both single learner NB and bagging NB perform better than the other prediction models, followed by the simple learner DStrump and bagging MLP. The group of learners, bagging DTable, bagging RepTree, RepTree, OneR and bagging OneR are the worst prediction models for the breastc data-set based on the performance measure *G-mean* and *TPR*. Even though the performance of *Accuracy* seems reasonably good for all the prediction models, it does not present the accuracy of the minority class. Because the performance of accuracy is influenced by the *TNR*, this observation is consistent with the existing research.

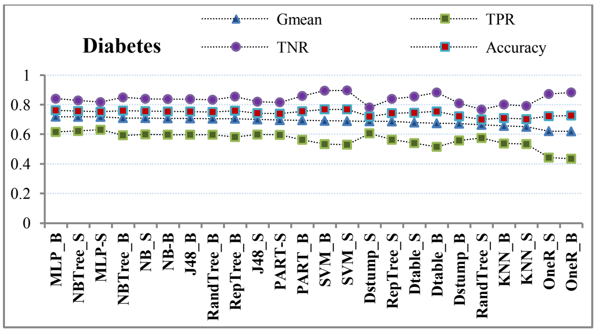


Figure 7: Comparison of the performance of prediction models on diabetes data-set.

Figure 7 presents the comparison of the performance of the prediction models on the diabetes data-set. The group of bagging MLP, NBTree, MLP and bagging NBTree are the best prediction models on this data-set, followed by NB and Bagging NB; while the group of learners, Bagging KNN, KNN, OneR and bagging OneR are the worst prediction models on this data-set.

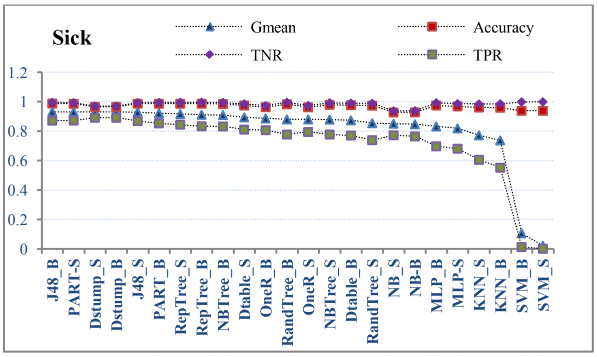


Figure8: Comparison of the performance of prediction models on sick data-set.

Figure 8 presents a comparison of the performance of the prediction models on the extremely imbalanced data-set, sick. We observe that *Accuracy* and *TNR* perform well for all the prediction models, because *Accuracy* is influenced by the *TNR* on this extremely imbalanced data-set. However, regarding the performance measures, *TPR* and *G-mean* of the accuracy of both the majority class and minority class, we observe that bagging J48 and PART perform best, followed by single DStump, bagging DStump, J48 and bagging PART, while the group learners, bagging KNN and SVM, and their single learners are the worst prediction models for this medical data-set.

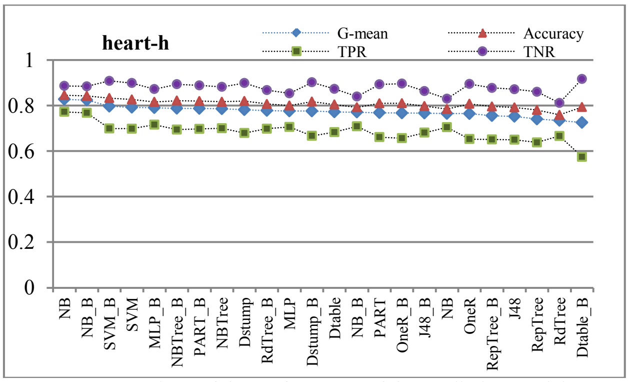


Figure 9: Comparison of the performance of the prediction models, bagging predictors and single learners on heart-h data-set.

Figure 9 presents a comparison of the performance of prediction models on the almost balanced heart-h data-set. Most prediction models perform well on this data-set, except the group of weak learners DStump and its bagging predictors. The group of learners, NB, bagging NB and SVMP, and bagging SVM are the best prediction models on this data-set.

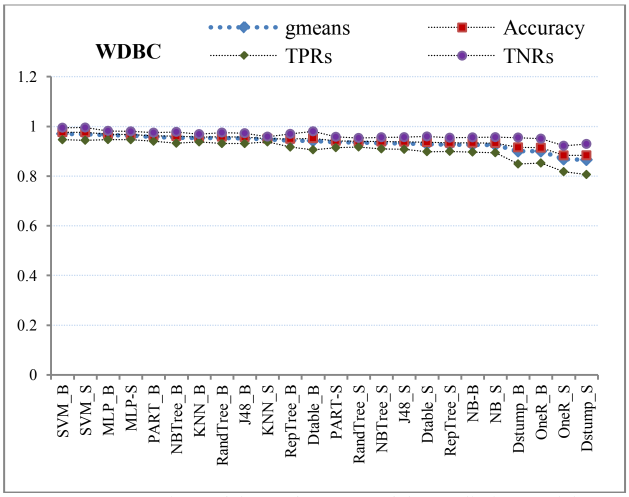


Figure 10: Comparison of the performance of the prediction models, bagging predictors and single learners on WDBC data-set.

Figure 10 presents a comparison of the performance of prediction models on the moderately imbalanced WDBC data-set. Most prediction models perform well on this data-set, except the group of weak learners, OneR, DStump, and their bagging predictors. The group of learners, bagging SVM, SVM, bagging MLP and MLP are the best prediction models on this data-set.

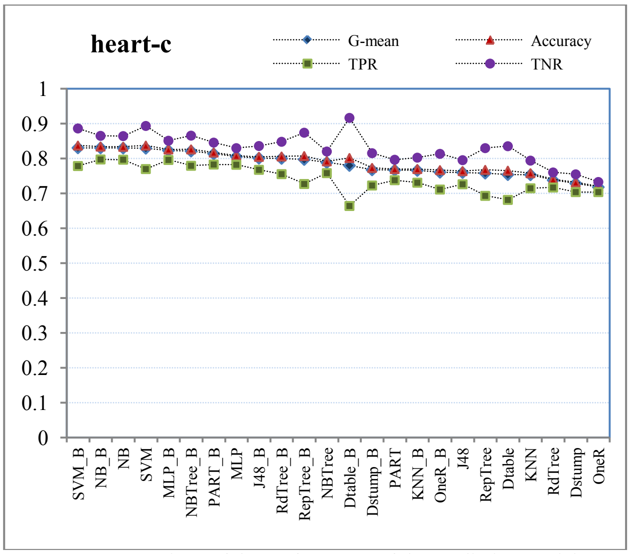


Figure 11: Comparison of the performance of the prediction models, bagging predictors and single learners on heart-c data-set.

Figure 11 presents a comparison of the performance of prediction models on the almost balanced heart-c data-set. The group of learners, bagging SVM, bagging NB, and NB are the best prediction models on this data-set, followed by SVM and bagging MLP; while the group of learners, KNN, Dstump and OneR are the worst prediction models on this data-set.

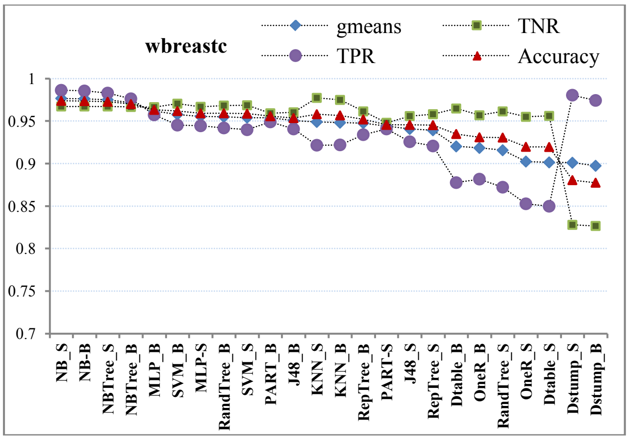


Figure 12: Comparison of the performance of the prediction models, bagging predictors and single learners on wbreastc data-set.

Figure 12 presents a comparison of the performance of the prediction models on the wbreastc data-set. The group of learners, NB, bagging NB, NBTree, and bagging NBTree are the best prediction models on this data-set, followed by bagging MLP and Bagging SVM; while the group of learners, Dstump and bagging Dstump are the worst prediction models on this data-set.

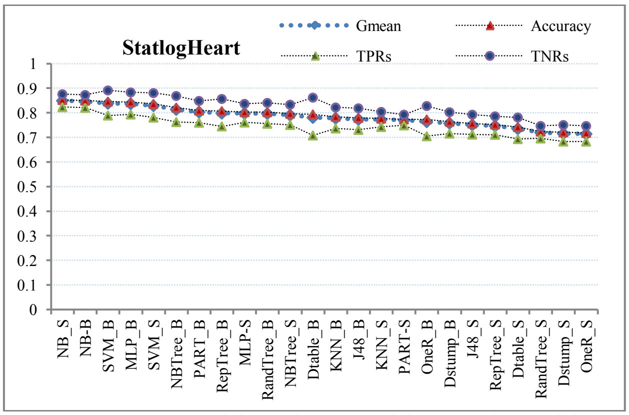


Figure 13: Comparison of the performance of the prediction models, bagging predictors and single learners on StatlogHert data-set.

Figure 13 presents a comparison of the performance of prediction models on the almost balanced StatlogHeart data-set. The group of learners NB and bagging NB are the best prediction models on this data-set, followed by bagging SVM and bagging SVM; while the group of learners, randTree, Dstump and OneR are the worst prediction models on this data-set.

Table 7: Best performance model for the natural class distribution on each individual data-sets.

	Best performance Model					P%
Name	G-mean	Err	TPR	TNR	Learners	
Heart-h	0.8239	0.1578	0.7679	0.8840	NB	0.36
Heart-c	0.831	0.1624	0.779	0.8867	SVM_B	0.45
stalogHeart	0.8492	0.1474	0.8233	0.876	NB	0.44
WDBC	0.97	0.0236	0.9462	0.9944	SVM_B	0.37
diabetes	0.7188	0.2384	0.6153	0.84	MLP_B	0.34
wbreastc	0.9767	0.0262	0.9672	0.9863	NB	0.34
breastc	0.6142	0.2703	0.4435	0.8507	NB	0.29
sick	0.932	0.0117	0.9959	0.8723	J48_B	0.06

Table 7 reports the best performance of prediction models for the natural class distribution on each individual medical data-set. Bagging predictors, SVM, MLP and J48 are the best prediction models for heart-c, WDBC, diabetes, and sick data-sets, respectively, while single learner NB is the best prediction models for heart-h, stalogHeart, wbreastc and breastc data-sets.

7.4 Comparison of the Performance of Bagging between the Natural Class Distribution and the Altered Class Distribution by Using Sampling Techniques on Each Medical Data-Set

In this subsection we report the performance of bagging predictors between the natural class distribution and the best achieved results by using sampling techniques on each medical data-set.

Tables 7 and 8 present the comparison of the performance of the bagging predictors between natural class distribution and the best achieved results by using sampling techniques on four medical data-sets. The first column indicates the name of a medical data-set and the bagging predictors; the second column presents the results from the natural class distribution which include *TPR, TNR and G-mean* of the accuracy rate on majority class and minority class; the third column presents the best achieved results based on *G-mean* by using sampling techniques which include *G-mean, TPR, TNR* and the proportion of the positive instances (*P%*) which refers to the level of the altered class distribution when bagging achieves the best performance on the *G-mean* measure. We also note that if the proportion of positive instances increases, the *TPR* will also increase but the *G-mean* may reduce.

The experimental results in the third column indicate the best achieved bagging performance based on the G-mean measure: the level of the class distribution is mostly about 50% on breastc, heart-c, and statlogHeart data-sets. This finding is consistent with previous research. However, the levels of class distribution are mostly 40% on WDBC and heart-h data-sets, and 30% on sick data-set, respectively, when the best bagging performance on the *G-mean* measure is achieved. In addition, there are interesting findings on both WDBC and sick data-sets in that when bagging NB achieves the best performance on the *G-mean* measure, the level of class distributions are 10% and 20% highlighted, respectively. This finding may be inconsistent with existing research, which assumes that traditional learning algorithms will perform better in a balanced situation than in an imbalanced situation.

The experimental results demonstrate that the sampling techniques can improve the performance of bagging predictors on the *G-mean* of the accuracy on the majority class and minority class over most medical data-sets, except for bagging OneR on the breastc data-set whose result is marked in red. The bagging performance on the TPR and TNR measures also improved at the same level of class distribution, except for NB on heart-h data-set with TNR measure marked as red.

.

breastc	**Natural Class Distribution**			**Sampling**			
Bagging	TPR	TNR	G-mean	G-mean	TPR	TNR	P%
J48	0.247	0.941	0.481	0.724	0.717	0.737	50%
RepTree	0.138	0.937	0.356	0.678	0.651	0.709	50%
RandTree	0.292	0.867	0.503	0.796	0.837	0.580	40%
NB	0.428	0.854	0.605	0.675	0.644	0.709	50%
SVM	0.308	0.865	0.516	0.690	0.695	0.685	50%
DStump	0.355	0.868	0.552	0.630	0.486	0.824	50%
OneR	0.207	0.904	0.431	0.619	0.505	0.769	50%
DTable	0.037	0.993	0.186	0.662	0.527	0.835	50%
PART	0.326	0.877	0.534	0.746	0.760	0.733	50%
KNN	0.338	0.887	0.546	0.802	0.782	0.822	50%
NBTree	0.284	0.915	0.509	0.731	0.732	0.732	50%
MLP	0.406	0.841	0.584	0.790	0.682	0.916	70%
heart-c	Natural			Sampling			
Bagging	TPR	TNR	G-mean	G-mean	TPR	TNR	P%
J48	0.7681	0.8364	0.8013	0.8872	0.8807	0.8941	50%
RepTree	0.7275	0.8745	0.7976	0.8496	0.82	0.8807	50%
RandTree	0.7558	0.8485	0.8007	0.9128	0.9052	0.9207	50%
NB	0.7978	0.8655	0.8309	0.8404	0.8019	0.8815	40%
SVM	0.779	0.8867	0.831	0.8461	0.8833	0.8109	60%
DStump	0.7232	0.8158	0.7679	0.7778	0.7454	0.8123	40%
OneR	0.7116	0.8139	0.761	0.7642	0.7407	0.7889	50%
DTable	0.6645	0.917	0.7804	0.8471	0.7904	0.9081	50%
PART	0.7833	0.8461	0.8139	0.9061	0.8956	0.917	50%
KNN	0.7312	0.803	0.7662	0.8983	0.9015	0.8956	50%
NBTree	0.7797	0.8661	0.8217	0.904	0.88	0.9289	50%
MLP	0.7964	0.8515	0.8234	0.9091	0.9074	0.9111	50%
Statlog Heart	Natural			Sampling			
Bagging	TPR	TNR	G-mean	G-mean	TPR	TNR	P%
J48	0.731	0.819	0.773	0.870	0.878	0.863	50%
RepTree	0.745	0.857	0.799	0.860	0.859	0.862	50%
RandTree	0.756	0.841	0.797	0.900	0.895	0.905	50%
NB	0.821	0.873	0.846	0.854	0.844	0.865	50%
SVM	0.789	0.891	0.839	0.865	0.853	0.877	50%
DStump	0.716	0.803	0.758	0.780	0.716	0.854	30%
OneR	0.705	0.828	0.764	**0.740**	0.726	0.756	50%
DTable	0.708	0.862	0.781	0.836	0.872	0.803	50%
PART	0.760	0.849	0.803	0.892	0.855	0.931	40%
KNN	0.737	0.822	0.778	0.891	0.863	0.919	40%
NBTree	0.763	0.869	0.814	0.900	0.856	0.946	40%
MLP	0.793	0.883	0.837	0.912	0.883	0.941	40%
heart-h	Natural			Sampling			
Bagging	TPR	TNR	G-mean	G-mean	TPR	TNR	P%
J48	0.6811	0.8633	0.7668	0.8548	0.8543	0.8562	50%
RepTree	0.6509	0.8771	0.7555	0.8282	0.8814	0.7794	60%
RandTree	0.6981	0.8676	0.7781	0.9035	0.8731	0.9353	40%
NB	0.7679	0.884	0.8239	0.8369	0.8076	0.8676	50%
SVM	0.6991	0.908	0.7966	0.8354	0.8794	0.7941	60%
DStump	0.667	0.9021	0.7757	0.7963	0.7412	0.8588	60%
OneR	0.6566	0.8963	0.7671	0.7944	0.801	0.7897	60%
DTable	0.5745	0.9165	0.7254	0.8297	0.8048	0.8571	50%
PART	0.6972	0.8883	0.7869	0.8665	0.8279	0.9077	40%
KNN	0.7085	0.8394	0.7711	0.8951	0.8731	0.9179	40%
NBTree	0.6943	0.8936	0.7876	0.8728	0.8346	0.9135	40%
MLP	0.716	0.8723	0.7903	0.8927	0.8596	0.9276	40%

Table 7: Compare the performance of bagging predictors on the G-mean measure between the natural class distribution and the altered class distribution by using sampling techniques on four data-sets: breastc, heart-c, statlogHeart and heart-h.

diabetes	Natural			Sampling			
Bagging	TPR	TNR	G-mean	G-mean	TPR	TNR	P
J48	0.597	0.837	0.707	0.861	0.834	0.890	40%
RepTree	0.581	0.855	0.705	0.824	0.780	0.871	40%
RandTree	0.597	0.833	0.705	0.876	0.844	0.909	40%
NB	0.597	0.837	0.707	0.726	0.740	0.712	60%
SVM	0.534	0.894	0.691	0.741	0.700	0.785	50%
DStump	0.558	0.809	0.672	0.696	0.620	0.792	40%
OneR	0.435	0.883	0.620	0.719	0.723	0.716	50%
DTable	0.515	0.882	0.674	0.776	0.778	0.774	50%
PART	0.563	0.859	0.695	0.852	0.824	0.881	40%
KNN	0.538	0.801	0.656	0.848	0.816	0.881	40%
NBTree	0.593	0.849	0.710	0.840	0.803	0.879	40%
MLP	0.615	0.840	0.719	0.812	0.833	0.793	50%
sick	Natural			Sampling			
Bagging	TPR	TNR	G-mean	G-mean	TPR	TNR	P
J48	0.872	0.996	0.932	0.973	0.967	0.979	30%
RepTree	0.834	0.997	0.912	0.965	0.954	0.976	30%
RandTree	0.778	0.997	0.881	0.972	0.964	0.980	40%
NB	0.765	0.939	0.848	0.880	0.864	0.898	**20%**
SVM	0.013	0.999	0.107	0.892	0.857	0.930	30%
DStump	0.892	0.970	0.930	0.934	0.896	0.974	70%
OneR	0.807	0.974	0.887	0.934	0.898	0.971	30%
DTable	0.771	0.991	0.874	0.941	0.902	0.982	30%
PART	0.854	0.995	0.922	0.973	0.967	0.979	30%
KNN	0.552	0.986	0.738	0.912	0.908	0.915	40%
NBTree	0.833	0.995	0.910	0.974	0.964	0.984	30%
MLP	0.698	0.993	0.832	0.951	0.964	0.938	50%
WDBC	Natural			Sampling			
Bagging	TPR	TNR	G-mean	G-mean	TPR	TNR	P
J48	0.931	0.972	0.951	0.974	0.968	0.981	40%
RepTree	0.917	0.970	0.943	0.968	0.959	0.978	40%
RandTree	0.931	0.975	0.953	0.979	0.971	0.986	40%
NB	0.897	0.956	0.926	0.937	0.905	0.970	**10%**
SVM	0.946	0.994	0.970	0.977	0.966	0.987	50%
DStump	0.849	0.955	0.900	0.925	0.940	0.910	70%
OneR	0.852	0.950	0.900	0.929	0.900	0.959	40%
DTable	0.906	0.980	0.942	0.958	0.961	0.954	40%
PART	0.940	0.975	0.957	0.979	0.977	0.981	40%
KNN	0.937	0.969	0.953	0.980	0.980	0.981	50%
NBTree	0.932	0.978	0.955	0.979	0.976	0.981	50%
MLP	0.947	0.982	0.964	0.979	0.972	0.985	50%
wbreastc	Natural			Sampling			
Bagging	TPR	TNR	G-mean	G-mean	TPR	TNR	P
J48	0.941	0.960	0.950	0.964	0.967	0.961	40%
RepTree	0.934	0.961	0.948	0.961	0.964	0.958	50%
RandTree	0.942	0.968	0.955	0.982	0.983	0.981	40%
NB	0.986	0.967	0.976	0.981	0.985	0.976	60%
SVM	0.945	0.971	0.958	0.979	0.988	0.970	80%
DStump	0.974	0.827	0.897	0.908	0.981	0.840	30%
OneR	0.882	0.957	0.918	0.935	0.962	0.908	60%
DTable	0.878	0.965	0.920	0.960	0.981	0.940	60%
PART	0.949	0.959	0.954	0.964	0.969	0.958	40%
KNN	0.922	0.975	0.948	0.981	0.982	0.981	60%
NBTree	0.976	0.967	0.972	0.983	0.983	0.984	30%
MLP	0.957	0.966	0.962	0.979	0.990	0.968	60%

Table 8: Compare the performance of bagging predictors on *G-mean* measure between the natural class distribution and the altered class distribution by using sampling techniques on four data-sets: diabetes, sick, WDBC and wbreastc.

8 Conclusions

This research investigates the performance of bagging predictors with respect to 12 different learning algorithms on 8 medical data-sets. We address the imbalance class distribution and unequal cost of mis-classification errors issues on medical data which may have high accuracy but poor performance on the TPR of minority class. We report the best performance prediction model for the natural class distribution on each individual medical data-set by comparing 12 single learners and 12 bagging predictors. In addition, we utilize sampling techniques to alter the class distribution at different imbalanced levels, and report the comparison of the bagging performance between the natural class distribution and the best achieved performance based on the *G-mean* measure at a certain level of class distribution. We note that by using sampling techniques to improve the performance of the bagging predictors, the level of the class distribution is mostly at 50% balanced level for three data-sets, breastc, heart-c, and statlogHeart; however, it is mostly at 40% for the diabetes, WDBC and heart-h data-set, and at 30% for the sick data-set. In addition, we also observe that the levels of class distribution for bagging NB to achieve the best performance on the *G-mean* measure are at 10% for the WDBC data-set and 20% for the sick data-set.

We investigated the effectiveness of bagging by using statistical tests. We also compared the performance of 12 bagging predictors on each of the medical data-sets; we observed that different bagging predictors behave differently for different medical data-sets. Bagging MLP performs well on most of these medical data-sets, except for the extremely imbalanced class distribution and high dimensional attributes large data-set ‘sick’; Bagging NB has the best performance on 4 out of 8 medical data-sets but performs poorly on two medical data-sets: sick and WDBC; Bagging J48 and Dstump perform well on the extremely imbalanced and high dimensional large data-set, sick. The full comparison of the performance of bagging predictors would allow data mining practitioners to choose proper learners and to understand what to expect when using bagging predictors for medical imbalanced applications.

References:

Bauer, E. & Kohavi, R. (1999) An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning,* 36(1): 105-139.

Breiman, L. (1996) Bagging predictors. *Machine Learning,* 24(2): 123-140.

Bunkhumpornpat, C., Sinapiromsaran, K. & Lursinsap, C. (2009): Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. *Proc.* PAKDD 2009 Conference, 475-482,

Chawla, N. V. (2010) Data mining for imbalanced datasets: An overview. *Data Mining and Knowledge Discovery Handbook,* 875-886.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002) Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research,* 16(1): 321-357.

Chawla, N. V., Lazarevic, A., Hall, L. O. & Bowyer, K. W. (2003) Smoteboost: Improving prediction of the minority class in boosting. *Knowledge Discovery in Databases: PKDD 2003,* 107-119.

Demšar, J. (2006) Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research,* 7(1-30.

Dietterich, T. (2000) An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning,* 40(2): 139-157.

Fawcett, T. (2006) An introduction to roc analysis. *Pattern Recognition Letters,* 27(8): 861-874.

Guo, X., Yin, Y., Dong, C., Yang, G. & Zhou, G. (2008) On the class imbalance problem. *Fourth International Conference on Natural Computation.* IEEE.

Han, H., Wang, W. Y. & Mao, B. H. (2005) Borderline-smote: A new over-sampling method in imbalanced data sets learning. *Advances in Intelligent Computing,* 878-887.

He, H. & Garcia, A. E. (2009) Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering,* 21(9): 1263-1284.

Kim, M.-J. & Kang, D.-K. (2010) Ensemble with neural networks for bankruptcy prediction. *Expert Systems with Applications,* 37(4): 3373-3379.

Liang, G. & Zhang, C. (2011): An empirical evalutation of bagging with different learning algorithms on imbalanced data. *Proc.* 7th International Conference on Advanced Data Mining and Applications Conference, 17th-19th December, Beijing, China, Springer.

Liang, G., Zhu, X. & Zhang, C. (2011a): An empirical study of bagging predictors for different learning algorithms. *Proc.* 25th AAAI Conference on Artificial Intelligence, AAAI 2011 Conference, 7-11 August, San Francisco, USA, 1802-1803, AAAI Press.

Liang, G., Zhu, X. & Zhang, C. (2011b): An empirical study of bagging predictors for imbalanced data with different levels of class distribution. *Proc.* 24th Australasian Joint Conference on Artificial Intelligence, AI 2011 Conference, 5th-8th December, Perth, Australia, Springer.

Liu, W. & Chawla, S. (2011): Class confidence weighted knn algorithms for imbalanced data sets. *Proc.* PAKDD 2011 Conference, 6635:345-356, Springer Berlin / Heidelberg.

Lopes, L., Scalabrin, E. & Fernandes, P. (2008) An empirical study of combined classifiers for knowledge discovery on medical data bases. *Advanced Web and NetworkTechnologies, and Applications,* 110-121.

Maimon, O., Rokach, L. & Chawla, N. V. (2010) Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook.* Springer US.

Maloof, M. (2003) Learning when data sets are imbalanced and when costs are unequal and unknown. *ICML-2003 workshop on learning from imbalanced data sets II.* Washington, DC.

Merz, C. & Murphy, P. (2006) *Uci repository of machine learning databases.*

Ng, W. & Dash, M. (2006) An evaluation of progressive sampling for imbalanced data sets. *Sixth IEEE International Conference on Data Mining Workshops.* IEEE.

Opitz, D. & Maclin, R. (1999) Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research,* 11(1): 169-198.

Phua, C., Alahakoon, D. & Lee, V. (2004) Minority report in fraud detection: Classification of skewed data. *ACM SIGKDD Explorations Newsletter,* 6(1): 50-59.

Provost, F. & Fawcett, T. (2001) Robust classification for imprecise environments. *Machine Learning,* 42(3): 203-231.

Provost, F., Fawcett, T. & Kohavi, R. (1998) The case against accuracy estimation for comparing induction algorithms. Citeseer.

Quinlan, J. (1996) Bagging, boosting, and c4.5. *Proceedings of the National Conference on Artificial Intelligence.*

Quinlan, J. R. (1986) Induction of decision trees. *Machine learning,* 1(1): 81-106.

Su, C. T. & Hsiao, Y. H. (2007) An evaluation of the robustness of mts for imbalanced data. *IEEE Transactions on Knowledge and Data Engineering,* 1321-1332.

Weiss, G. M. & Provost, F. (2003) Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research,* 19(1): 315-354.

West, D., Dellana, S. & Qian, J. (2005) Neural network ensemble strategies for financial decision applications. *Computers & Operations Research,* 32(10): 2543-2559.

Witten, I. H. & Frank, E. (2005) *Data mining: Practical machine learning tools and techniques,* San Francisco, Morgan Kaufmann.