

An Effective Approach for Imbalanced Classification: Unevenly Balanced Bagging

Guohua Liang¹ and Anthony G Cohn^{2,1}

The Centre for Quantum Computation & Intelligent Systems, FEIT,
University of Technology, Sydney NSW 2007, Australia¹
School of Computing, University of Leeds, Leeds LS2 9JT, UK²
Guohua.Liang@student.uts.edu.au;a.g.cohn@leeds.ac.uk

Abstract

Learning from imbalanced data is an important problem in data mining research. Much research has addressed the problem of imbalanced data by using sampling methods to generate an equally balanced training set to improve the performance of the prediction models, but it is unclear what ratio of class distribution is best for training a prediction model. Bagging is one of the most popular and effective ensemble learning methods for improving the performance of prediction models; however, there is a major drawback on extremely imbalanced data-sets. It is unclear under which conditions bagging is outperformed by other sampling schemes in terms of imbalanced classification. These issues motivate us to propose a novel approach, unevenly balanced bagging (*UBagging*), to boost the performance of the prediction model for imbalanced binary classification. Our experimental results demonstrate that *UBagging* is effective and statistically significantly superior to single learner decision trees *J48* (Single.*J48*), bagging, and equally balanced bagging (BBagging) on 32 imbalanced data-sets.

Introduction

Imbalanced class distribution (Weiss and Provost 2003) refers to a situation in which the numbers of training samples are unevenly distributed among different classes. The imbalanced class distribution problem is an important challenging problem in data mining research. Bagging (Breiman 1996) is an effective ensemble method to improve the performance of the prediction model. However, in an extremely imbalanced situation, bagging performs poorly in rendering predictions of the minority class. This is the major drawback of bagging when dealing with an imbalanced data-set.

Sampling techniques are considered to be an effective way to tackle the imbalanced class distribution problem. Goebel states that there must be situations in which bagging is outperformed by other sampling schemes in terms of predictive performance (Goebel 2004). We believe that in extremely imbalanced situation, bagging can be outperformed by other sampling schemes. These issues motivate us to propose a new sampling scheme,

unevenly balanced bagging (*UBagging*), for outperforming the bagging prediction models on imbalanced data-sets.

Most research on existing bagging-based sampling schemes for imbalanced data, e.g. (Li 2007; Hido, Kashima, and Takahashi 2009), focused on using sampling methods to provide a set of equally balanced or average-balanced training sub-sets for training classifiers to improve the performance of the prediction models for imbalanced classification. (Liang, Zhu, and Zhang 2011; 2012) investigated the impact of varying the degree of class distribution from 10% to 90% ($|P_i| : |P_i| + |N_i|$) with the same bagged size in a set of training sub-sets in each ensemble learning. To our knowledge, nobody has used a set of training sub-sets with both different bag sizes and varying ratios of class distribution in the ensemble as a sampling scheme to try to outperform bagging for imbalanced data.

This paper proposes the *UBagging* approach, a new sampling scheme to generate a set of unevenly balanced bootstrap samples to form a set of training sub-sets in an ensemble to boost the performance of the prediction model on imbalanced data-sets. The key contributions of this approach are as follows. (1) A new sampling scheme, *UBagging*, is proposed. (2) Empirical investigation and statistical analysis of the performance of the four prediction models, Single.*J48*, bagging, BBagging and *UBagging* are comprehensively performed. (3) Our *UBagging* approach is demonstrated to be effective and statistically significantly superior to the other three prediction models at a 95% confidence interval on 32 imbalanced data-sets.

The UBagging Algorithm

Algorithm 1 outlines our new approach. Our designed framework is very different from previous approaches for imbalanced classification. In each sub-set of the training set, the positive instances are randomly selected with replacement from the entire positive class, where the number of positive instances $|P_i|$ have the same size as the entire positive class, $|P|$; the negative instances are randomly selected from the negative class of the original training data with replacement, where the number of negative instances $|N_i|$ is incrementally increased by 5% of $|P|$ from $\frac{1}{2} * |P|$ to $2 * |P|$. As a result, the size and class distribution of the sub-sets are different in each of the 31 bags in the ensemble.

Algorithm 1: Unevenly Balanced Bagging

Input:

D , original training set, containing $|P|$ positive and $|N|$ negative instances;
a learning scheme, eg. $J48$;

Output: A composite model, C^* .

Method:**Do**

Create unevenly balanced bootstrap samples of size $|D_i|$ sub-sets, $D_i = P_i + N_i$ where,
 P_i and N_i are randomly drawn with replacement from P and N , respectively, where:
 $|P_i| = |P|$ and;
 $|N_i| = (0.5 + 0.05 * i) * |P|$;
Train each base classifier model C_i from D_i ;

while $|N_i| < 2 * |P|$

To use the composite model, C^* for a test set T on an instance x where its 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.

Experimental Results and Analysis

This section presents the experimental results and analysis, comparing the performance of the prediction models based on two evaluation metrics, F_{value} and G_{mean} . A 10-trial 10-fold cross-validation evaluation is employed for this study. The $J48$ with default parameters from WEKA is used as the base learner.

Table 1: Comparison of the performance of four prediction models based on F_{value} and G_{mean}

Evaluation Methods	F_{value}				G_{mean}			
	SingleJ48	Bagging	Bbagging	UBagging	SingleJ48	Bagging	Bbagging	UBagging
Average	0.656	0.687	0.772	0.787	0.711	0.739	0.888	0.902
STD	0.284	0.276	0.207	0.202	0.274	0.254	0.087	0.076
Average Rank	3.64	2.77	2.37	1.22	3.8	3.14	2.05	1.02
“Critical Difference”	0.829				0.829			

Table 1 presents the summary of the experimental results, which respectively indicate the average of the evaluation metrics with standard deviation (STD) and the average rank of evaluation metrics with “Critical Difference” of the Nemenyi test over 32 data-sets taken from (Merz and Murphy 2006). The results indicate that $UBagging$ performs the best on average with the smallest STD and average rank based on both evaluation metrics, F_{value} and G_{mean} , across all data-sets (results in bold indicate the best overall performance out of the four classifiers).

The Null Hypothesis of the Friedman test is rejected, so a post-hoc Nemenyi test is required to calculate the “Critical Difference” to determine and identify where one prediction model is significantly different from another (Demšar 2006).

Figure 1 presents a comparison of the performance of the prediction models with the Nemenyi test, where the x -axis indicates the average rank of F_{value} and G_{mean} , respectively, the y -axis indicates the ranking order of the four prediction models, and the horizontal bars indicate the “Critical Difference”. If the horizontal bars between prediction models do not overlap, it means there is a

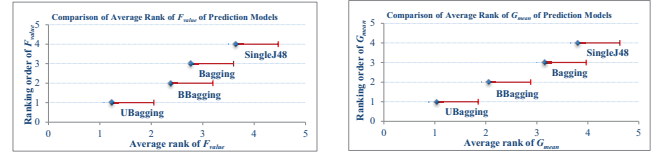


Figure 1: Comparison of the performance of four prediction models with the Nemenyi test, where the x -axis indicates the average rank of F_{value} and G_{mean} , respectively, the y -axis indicates the ranking order of the four prediction models, and the vertical bars indicate the “Critical Difference”.

statistically significant difference between the prediction models at a 95% confidence interval. The results indicate that based on F_{value} and G_{mean} , our proposed $UBagging$ is statistically superior to the other three prediction models.

Conclusion

This paper proposes a new $UBagging$ approach to boost the performance of the prediction model for imbalanced binary classification. This approach is different from previous approaches, which to the best of our knowledge all use identically sized bags (or nearly identical) to improve the performance of the bagging predictor to solve imbalanced classification problems.

The experimental results demonstrate that our new $UBagging$ approach is statistically significantly superior to the other three prediction models at a 95% confidence interval on two evaluation metrics over 32 imbalanced data-sets. We believe the success of these results will also apply to other base learners, and initial experiments with an SVM indicate support for this hypothesis.

References

- Breiman, L. 1996. Bagging predictors. *Machine Learning* 24(2):123–140.
- Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* 7:1–30.
- Goebel, M. 2004. *Ensemble learning by data resampling*. Ph.D. Dissertation, University of Auckland, NZ.
- Hido, S., Kashima, H., and Takahashi, Y. 2009. Roughly balanced bagging for imbalanced data. *Statistical Analysis and Data Mining* 2(5-6):412–426.
- Li, C. 2007. Classifying imbalanced data using a bagging ensemble variation (BEV). In *Proceedings of the 45th ACM Annual Southeast Regional Conference*, 203–208.
- Liang, G., Zhu, X., and Zhang, C. 2011. An empirical study of bagging predictors for imbalanced data with different levels of class distribution. In *Proceedings of the 24th Australasian Conference on Artificial Intelligence*, 213–222.
- Liang, G., Zhu, X., and Zhang, C. 2012. The effect of varying levels of class distribution on bagging for different algorithms: An empirical study. *International Journal of Machine Learning and Cybernetics*. <http://link.springer.com/article/10.1007%2Fs13042-012-0125-5>.
- Merz, C., and Murphy, P. 2006. UCI repository of machine learning databases. <http://archive.ics.uci.edu/ml/>.
- Weiss, G., and 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.