Volume 5, Issue 6, June 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Learning from Imbalanced Data of Diverse Strategies with Investigation

Mohammad Imran[1]
Regd No: PP.COMP.SCI&ENG.0308C,
Research Scholar, Computer Science and Engineering,
Rayalaseema University, Kurnool-518007,
Andhra Pradesh, India

Dr.Vaddi Srinivasa Rao[2]
Professor & Head
Department of CSE, Velagapudi Ramakrishna
Siddhartha Engineering College, Vijayawada–520007,
Andhra Pradesh, India

A.Vijaya Kumar[3]
Associate Professor, IT Nalla Malla Reddy
Engineering College, Ranga Reddy,
Telangana, India

Mohammed Afroze[4]
Assistant Professor, CSE Muffakham Jah College
of Engineering and Technology, Banjara Hills,
Hyderabad-500034, Telangana, India

Dr.Ahmed Abdul Moiz Qyser[5]
Professor and Head Department of CSE,
Muffakham Jah College of Engineering and
Technology, Banjara Hills, Hyderabad-500034,
Telangana, India

Abstract: A Learning from imbalanced data has no predictable conclusions from existing studies in regards to the effectiveness of diverse methodologies. In this paper we apply bias-variance analysis to study the utility of different strategies for imbalanced learning. We conduct experiments on 15 real-world imbalanced datasets of applying various re-sampling and induction bias adjustment strategies to the standard naïve bayes, k-nearest neighbour (k-NN) and decision tree learning algorithms. Our main findings include: Imbalanced class distribution is primarily a high bias problem, which partly explains why it impedes the performance of many standard learning algorithms. Compared to the re-sampling strategies, adjusting induction bias can more significantly vary the bias and variance components of classification errors. Especially the inverse distance weighting strategy can significantly reduce the variance errors for k-NN. Based on these findings we offer practical advice on applying the re-sampling and induction bias adjustment strategies to improve imbalanced learning.

Keywords: Bias-Variance Analysis, Imbalanced, Learning

I. INTRODUCTION

In many applications class distribution is imbalanced, and class of minority is by far of the primary concern. In these applications, typically the purpose of classification learning is to accurately predict the class of minority. As an example predicting defects in source code is of uttermost importance in software development projects, but defects only occur at a modest ratio of 5-10%. Accurate prediction of software defects can significantly reduce costs for software development.Class imbalance has been reported to hamper the performance of standard classification models, whose aim is usually to optimize the overall accuracy.

For example, the standard decision tree model tends to be inundating by the class of majority and ignore the class of minority when making a decision about class labels. Re-sampling and adjusting induction biases have been popular approaches to combating class imbalance. Changing the prevalence of positive and negative examples by sampling is a widely used method for addressing class imbalance. Strategies include random under-sampling with class of majority, random over-sampling with class of minority and more advanced intelligent over-sampling techniques [1] [2]. Adjusting the induction bias to favour the minority class is another method to accomplish accurate taxonomy on the class of minority. A natural question is what effect these imbalanced learning strategies have on the behaviour of standard learning algorithms. In particular, how well a model fits the problem under consideration and to what extent a model is affected by variation in class distribution. To this end we employ the bias and variance analysis of classification errors [3] to improve our understanding of the behaviours of different learning algorithms in the presence of class imbalance, and the effectiveness of sampling and imbalance induction bias adjustment on different learning models. With the bias-variance decomposition, three types of classification errors are distinguished: the bias errors are the systematic errors associated with the learning algorithm and the problem domain, the variance errors are caused by variations in samples and the intrinsic errors are associated with the inherent uncertainty of the problem domain. Generally high bias errors indicate that a model is not correct for the problem domain and high variance errors indicate unstable classification by the model. An intrinsic error is associated with noise of the problem domain and is independent of the learning algorithm. We employ the bias and variance decomposition of classification errors to study the behaviour of three representative learning algorithms, the C4.5 decision tree algorithm [4], the naive bayes (NB) [5] [6] [7], and the k-nearest neighbour (k-NN). We also study how random under- and over-sampling and advanced sampling techniques (See Section 3) vary the bias and variance components of errors for learning algorithms. We conduct a large-scale empirical study on 15 imbalanced datasets from the UCI repository and other disciplines.

Our main findings include: Imbalanced class distribution impedes the performance of standard learning algorithms in general, but depending on the learning algorithm, having varying effects on the bias and variance components of errors. The re-sampling strategies have varying effects on the bias or variance of learning algorithms. On the other hand adjusting the induction bias can significantly reduce the bias or variance components of errors, depending on learning algorithms. Based on this analysis we offer practical advice on applying the various re-sampling and induction bias adjustment strategies to combat the imbalanced learning problem.

II. RELATED WORK

A few empirical studies have studied and compared different sampling techniques [8][9][10]. Though, no regular conclusions have drawn since these studies. Most of these studies use a few datasets for experiments and so the conclusion is hard to generalize. A large scale experimental study was conducted in [10], but datasets used in the study are not publicly available. It was found in that the effectiveness of re-sampling for imbalanced learning depends on the evaluation metrics and base learning algorithms. All these previous studies have examined the effectiveness of imbalanced learning strategies on classification accuracy. We concentrate on explaining the behaviour of imbalanced learning strategies with the bias and variance decomposition of classification errors. Our bias and variance analysis relates the in-consistent behaviour of re-sampling to that it does not generally have consistent effect on the bias or variance errors of learning algorithms. Importantly we offer practical advice on how to combine re-sampling strategies for effective imbalanced learning. When classification errors (misclassification costs) of different classes are distinguished, accuracy maximization is replaced with cost minimization {high cost is associated with misclassifying minority samples. Cost-sensitive learning methods [11] have been proposed to learn from imbalanced class distribution. In [12] the problem of optimal learning with different misclassification cost is studied. It is shown that in theory that rebalancing the positive and negative distribution has little effect on the decision tree and Bayesian methods. However this general theoretical result does not necessarily suggest that re-sampling strategies do not work in specific applications. In an excellent survey by Weiss [13], techniques for imbalanced learning were reviewed. Sampling and adjusting decision bias are recognised as a commonly used technique for dealing with rarity, but no conclusion was drawn regarding their effectiveness. With some recent developments, advanced sampling techniques were proposed [14] for specific imbalanced learning applications. However the general utility of these techniques are yet to be studied. The bias and variance analysis of classification errors [3] is a widely used approach to provide insight into the error performance of classifiers. It has been used in various studies to compare the relative performance of different learning models, for example [15] [16] [17]. Re-Sampling Strategies for Imbalanced Learning Based on the assumption that standard learning methods perform better with equal class distribution, re-sampling training instances has been proposed for imbalanced learning.

1. UNDER-SAMPLING WITH OVER-SAMPLING RANDOMLY

Randomly under-sampling with over-sampling training instances is two basic methods of re-sampling for imbalanced learning. With under-sampling, examples of the majority class are randomly eliminated so as to achieve balanced class distribution. With over-sampling, examples of the class of minority are randomly replicated to achieve even class distribution. In essence random over-sampling does not introduce new examples to directly bias the induction process.

Some studies have shown that, compared with under-sampling, simple over-sampling is less effective at improving recognition of the minority class [9]. However another study that used artificial domains came to the opposite conclusion [8].

1.1 SOPHISTICATED SAMPLING METHODS

A more sophisticated sampling method is to combine under-sampling and over-sampling to achieve balanced class distribution. This potentially can remedy the drawbacks when under-sampling and over sampling are used separately. The Synthetic Minority Oversampling Technique (SMOTE) [2] generates minority-class examples by adding examples from the line segments that join the k minority-class nearest neighbours. This presumably leads to better generalization compared with random over-sampling. It was shown that a combination of over-sampling the minority class using SMOTE and under-sampling the majority class can achieve better classifier performance than only under-sampling the majority class. However, the effect of SMOTE alone on imbalanced learning has not been extensively studied.

2. ADJUSTING INDUCTION BIAS FOR IMBALANCED LEARNING

In this section we discuss three popular base learning algorithms and where applicable, strategies adjusting their induction bias for imbalanced learning.

2.1 THE DECISION TREE

Some strategies have been proposed adjusting the decision tree induction to be more sensitive to imbalanced class distribution [10]:

- For imbalanced class distribution, pruning a decision tree can over generalize and completely ignore the positive class, and so decision trees are fully grown without pruning.

- Based on similar consideration, the minimal number instances for leaves of a decision tree are set to one rather than a number > 1.
- With Laplace smoothing (Good 1965) the probability for the positive class at a leaf node is estimated as $\frac{Lp+1}{Lp+Ln+2,}$ where L_p and L_n are correspondingly the amount of positive and negative samples at the leaf. It has been shown that Laplace smoothing improves the tree performance for skewed class distribution.

2.2 THE NAIVE BAYES

The Naive Bayes (NB) is a uncomplicated probabilistic induction model based on the theorem of Bayes [6]. NB estimates probabilities based on the attribute independence assumption. Although this assumption does not hold for many problems, NB of-ten exhibits competitive classification accuracy compared with other learning algorithms. NB has a very strong induction bias and does not have any parameters that can be adjusted for imbalanced class distribution.

TABLE 1: RELATIONSHIP OF K-NN, C4.5 AND NB WITH BIAS-VARIANCE

lgorithm	Induction bias	Correct		Incorrect	
		Bias error	Variance error	Bias error	Variance error
k-NN	Weak	low	high	high	high
C4.5	Medium	varying	varying	varying	varying
NB	Strong	low	low	high	low

2.3 THE K-NN (K-NEAREST NEIGHBOUR)

With the k-NN (k-nearest neighbour) [18] algorithm, class labels of the k training instances closest to a test instance help determine the class label of the test instance. Inverse distance weighting is to weigh the vote of each neighbour according to the inverse of its distance from the test instance [19]. By taking the weighted average of the k neighbours nearest to the test instance smoothes out the impact of isolated noisy training in-stances. Furthermore it lifts the weight of instances from the minority class closest to the test instance a point that has been largely overlooked by existing studies.

2.4 THE BIAS-VARIANCE INVESTIGATION

The bias-variance analysis of classification errors is a useful tool for analysing classifier behaviour. This investigation decomposes classification errors into terms of three, derived with suggestion to the performance of a learning algorithm when trained with different training sets drawn from some reference distribution of training sets:

- Squared bias denotes the systematic component of classification errors, how closely a learner describes the decision surfaces for a domain.
- Variance describes the component of classification errors from sampling, how sensitively a learner responds to variations in the training sample.
- Intrinsic noise measures the degree to which the target quantity is inherently unpredictable, which equals the expected cost of the Bayes optimal classifier.

There have been several proposals for the definition of the three terms for classification learning. The definition by Kohavi and Wolpert [3] is widely used and is the definition we will use in this study. Given that an error has cost 1 and a correct prediction has cost 0, the expected error rate for a target function f and a training dataset of size m is:

$$err = sum_x P(x)(noise_x^2 + bias_x^2 + variance_x)$$

where x ranges over the instance space, and P(x) is the prior probability of x. In practical experiments it is impossible to estimate the intrinsic noise. The algorithm proposed in [3] generates a bias term that includes the intrinsic noise. In their method, the training dataset is divided into a training pool and a test pool randomly. Each pool contains 50% of the training instances. Fifty training sets are generated from the training pool by random sampling. Classifiers are trained on each of the 50 training set, and bias and variance errors are estimated from the classifiers on the test set.

Generally there are bias-variance tradeoffs. When adjusting learning algorithm so that it is more sensitive to the training samples, its bias errors shrink but the variance errors increase. Learning models that over fit the given training data often have high variance errors their results depend closely on the given training data and thus vary for different training datasets. On the contrary learning models with a strong induction bias are less likely to over fit and bias is a source of prediction errors if the induction bias of the model is not correct for a domain.

General description of the C4.5 decision tree, k-nearest neighbour and Naive Bayes learning algorithms in terms of their effect on the bias and variance components of in

classification errors are presented in Table 1. With the strong attribute-value independence assumption during classification, Naive Bayes has a strong induction bias. If the induction bias of NB is correct for the problem domain, then NB demonstrates low bias errors otherwise high bias errors. Without any representation model, the classification decision of k-NN does not have induction bias and its classification errors mainly come from variations in the distribution of training data. With a decision tree as the representation model, C4.5 has a medium level of induction bias. As a result the classification errors of C4.5 can come from the bias, variance or both components.

TABLE 2. EXPERIMENTS OF DATASETS IN ORDERED WITH DIMINISHING LEVEL OF SKEWEDNESS.

ID	Dataset	#instances	#attr(numerical, nominal)	class (minority, majority)	Minority %
1	Oil*	990	47(47,0)	(true, false)	4.35
2	hypo-thyroid	3163	25(7,18)	(hypothyroid, negative)	4.77
3	PC1*	1109	21(21,0)	(true, false)	6.94
4	Glass	214	9(9,0)	(3, remainder)	7.94
5	Flag	194	28(10,18)	(white ,remainder)	8.76
6	Satimage	6435	36(36,0)	(4, remainder)	9.73
7	CM1*	498	21(21,0)	(true, false)	9.84
8	New-thyroid	215	5(5,0)	(3,remainder)	13.95
9	KC1*	2109	21(21,0)	(true, false)	15.46
10	SPECT	267	22(0,22)	(0,1)	20.6
11	Hepatitis	155	19(6,13)	(1,2)	20.65
12	Vehicle	846	18(18,0)	(van, remainder)	23.52
13	Splice-ei	3190	60(0,60)	(EI, remainder)	24.04
14	Haberman	306	3(3,0)	(2,1)	26.47
15	German	1000	20(7,13)	2,1)	30

We can now characterise performance of the three base learning algorithms for imbalanced learning in terms of bias-variance decomposition. We can also characterise the effect of various re-sampling and induction bias adjustment strategies on the bias and variance components of errors.

III. EXPERIMENT DESIGN

Our study will focus on the two-class problem with a minority (positive) class and a class of majority. We compile datasets from various sources to study the utility of re-sampling and induction bias adjustment strategies for classification. Fifteen real-world datasets from extremely imbalanced to reasonably imbalanced are used in our experiments, as listed in Table 2. UCI [20] imbalanced 2-class datasets include those from natural 2-class domains, and those constructed by choosing a minority class as the positive and the remainder as negative instances. The Oil dataset (Kubat et al. 1998) (marked with *) has been extensively used in imbalanced learning experiments. PC1, CM1 and KC1 (marked with *) contain metrics data at the module level for predicting defects in NASA software development projects (http://mdp.ivv.nasa. gov /index.html).

In our experiments, we use K-Means, NB and IBk of the Clementine [21] data mining software for the base algorithms C4.5 [4] decision tree, NB and k-NN. The base algorithms with default settings,

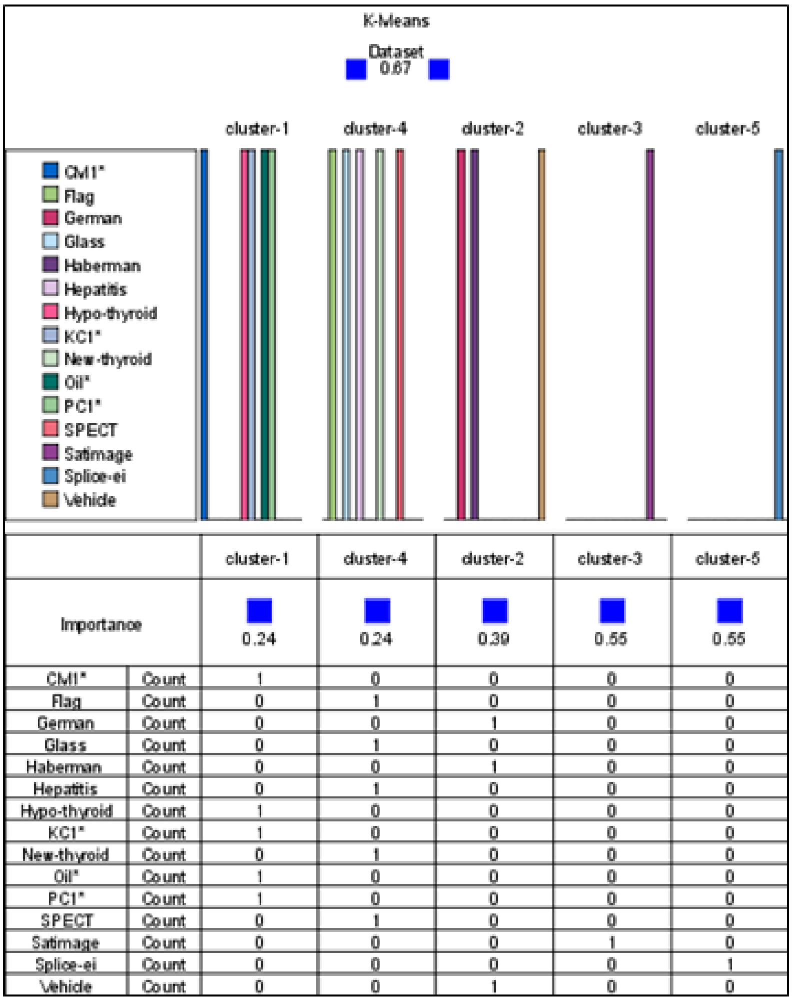
which usually are designed for uniform class distribution, are compared against their settings for adjusting induction bias for skewed class distribution. Specifically, for J48 the imbalance-favourable settings are without pruning, with Laplace-smoothing and that minimum one instance is allowed for a leaf node. For IBk, the imbalance-favourable settings are k=3 and inverse-distance weighted voting, There are not any parameter settings for adjusting bias for imbalanced distribution.We use the instance re-sampling filters in WEKA to implement the re-sampling strategies in Section 3. For under-sampling, the majority class is randomly under-sampled with replacement so that it has the same number of instances as the minority class. For over-sampling, the minority class is randomly over- sampled so that it has the same number of instances as the majority class. The SMOTE filter in WEKA is used for the SMOTE over-sampling strategy.

1. THE BIAS-VARIANCE ANALYSIS OF IMBALANCED LEARNING

In our experiments we employ the bias and variance decomposition software in the WEKA toolkit to estimate the squared bias and intrinsic noise combined error and the variance error component for classification algorithms. The bias and variance decomposition algorithm in WEKA precisely follows the approach of [3], as described in Section 4. The bias-variance decomposition for base learning algorithms Fig a. shows the bias and variance decomposition of expected errors for the base algorithms C4.5, k-NN and NB on 15 datasets in our experiments. Generally for all three base algorithms, the bias component is the dominant source of errors. Not surprisingly NB has the highest bias component of errors | except on Oil where bias comprises 43.94% of errors, on all other datasets bias is the bigger proportion of errors, com-prising on average 81.02% of errors. C4.5 and k-NN demonstrates varying bias-variance decomposition on 15 datasets, with the bias portion of errors ranging from 43.82% for C4.5, 51% for k-NN and 92.94% for NB. Our analysis suggests that imbalanced (C4.5 on Vehicle) to 98.10% (C4.5 on Flag).The BVD profile for base algorithms differ on each dataset. For example on the most imbalanced Oil dataset, the bias component of errors for C4.5, k-NN and NB are dramatically different, 57.89% for C4.5, 79.79% for k-NN and 43.94% for NB respectively. On the Vehicle dataset, the bias component is respectively 43.82%. class distribution has different effect on the base learning algorithms and it varies significantly for different problems. This complex profile of bias-variance component suggests that learning from imbalanced class distribution is a challenging problem.

1.1 THE BIAS-VARIANCE DECOMPOSITION FOR SAMPLING TECHNIQUES

A relatively large number of instances in the training dataset is needed to ensure accurate estimation of errors. In our experiments the smallest dataset (Hepatitis) contains 155 instances, which we consider sufficiently large. Under sampling the majority class to match the minority can result in some datasets have a very small number of instances We chose datasets whose total number of instances is at least 100 after under-sampling. From Fig. 1 it can be seen that generally random under-sampling increases both the bias and variance errors for all three base learning algorithms, and the increase in variance errors is more pronounced than that in the bias errors. k-NN demonstrates the most consistent and significant response to under-sampling on all 10 datasets both its bias and variance components of errors significantly increase, and the increase in variance is more pronounced. C4.5 is also very sensitive to under-sampling, and shows increment in both bias and variance errors on all 10 datasets. In contrast NB is not so sensitive to the under-sampling strategy.



		cluster-1	duster-4	duster-2	cluster-3	cluster-5
Importance		0.24	0.24	0.39	0.55	0.55
CMI*	Count	1	0	0	0	0
Flag	Count	0	1	0	0	0
German	Count	0	0	1	0	0
Glass	Count	0	1	0	0	0
Haberman	Count	0	0	1	0	0
Hepatitis	Count	0	1	0	0	0
Hypo-thyroid	Count	1	0	0	0	0
KC1*	Count	1	0	0	0	0
New-thyroid	Count	0	1	0	0	0
Oil*	Count	1	0	0	0	0
PC1*	Count	1	0	0	0	0
SPECT	Count	0	1	0	0	0
Satimage	Count	0	0	0	1	0
Splice-ei	Count	0	0	0	0	1
Vehide	Count	0	0	1	0	0

Fig1: Bias variance decomposition using k-means

On KC1 the bias errors for under-sampling remain the same as that of the standard NB.As a summary, our experiments have demonstrated that either random over-sampling or SMOTE intelligent over-sampling does not significantly change the bias errors of the three base learning algorithms. This can be explained by that the generated new samples are either replicates or near replicates of existing positive samples and they do not produce effect on the decision boundary for classes. Over sampling generally can also negatively affect the variance errors of the decision tree and k-nearest neighbour models, and it does not change the variance of NB. In contrast, under-sampling significantly changes the bias and variance of base algorithms, due to the fact that some “important" samples affecting the decision for class boundary may have been removed.

However, generally the effect is “negative” that is the bias and variance errors of all algorithms are ex-acerbated rather than reduced. Comparing the imbalance bias adjustment strategies of C4.5 and k-NN, the strategies in the decision tree algorithm has focused on modifying the representation tree for classification, especially towards the leaves at the bottom of the decision tree. Such strategies exacerbate the variance errors of the decision tree model. As a learning algorithm without explicit model representation, the imbalance induction bias adjustment of k-NN reduces both the bias and variance errors of the learning algorithm. This shows that the strategy has improved both the generality and stability of the k-NN algorithm.

IV. CONCLUSIONS

In this paper we have studied the re-sampling approach and the adjusting induction bias approach for employing standard learning algorithms for imbalanced classification. The re-sampling strategies we consider include random over-sampling, random under-sampling and SMOTE intelligent oversampling. We employ bias-variance analysis to study the behaviour of re-sampling and imbalance bias adjustment on 15 real-world imbalanced datasets for popular algorithms, including the decision tree, Naive Bayes and k-nearest neighbour. We have found that imbalanced class distribution impedes the performance of standard learning algorithms in general, but depending on the learning algorithm, having varying effects on the bias and variance components of errors. For the naive bayes algorithm, class imbalance mainly presents as a high bias problem, whereas for the decision tree and k-nearest neighbour models, errors can come from either the bias or variance component, depending on the application domain. More research is needed to investigate how to best combine under-sampling and over-sampling.

REFERENCES

[1] Kubat, M. & Matwin, S. (1997), Addressing the curse of imbalanced training sets: One sided selection, in `Proc. of 14th International Conference on Machine Learning', Morgan Kaufmann, Nashville, Tenesse, USA, pp. 179-186.

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002), `SMOTE: Synthetic minority oversampling technique', Journal of Artificial Intelligence Research 16, 321-357.

[3] Kohavi, R. & Wolpert, D. (1996), Bias plus variance decomposition for zero-one loss functions, in `Proc.ICML'.

[4] Quinlan, J. (1993), C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers.

[5] Good, I. (1965), The estimation of probabilities: An essay on modern bayesian methods, M.I.T. Press.

[6] Duda, R. & Hart, P. (1973), Pattern classi_cation and scene analysis, John Wiley and Sons, New York.

[7] Langley, P., Iba, W. & Thompson, K. (1992), Ananalysis of bayesian classi_ers, in `Proc. Tenth National Conference on Arti_cial Intelligence'.

[8] Japkowicz, N. & Stephen, S. (2002), `The class imbalance problem: a systematic study', Intelligent Data Analysis 6(5), 429-450.

[9] Drummond, C. & Holte, R. C. (2003), C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling, in `Workshop on Learning from Imbalanced Data Sets II'.

[10] Hulse, J. V., Khoshgoftaar, T. M. & Napolitano, A. (2007), Experimental perspective on learning from imbalanced data, in `Proc. Int'l Conference on Machine Learning'.

[11] Domingos, P. (1999), MetaCost: A general method for making classi_ers cost-sensitive, in `Proc. ACM SIGKDD', pp. 155-164.

[12] Elkan, C. (2001), The foundations of cost-sensitive learning, in `Proceedings of IJCAI 2001'.

[13] Weiss, G. M. (2004), `Mining with rarity: A unifying framework', SIGKDD Explorations 6(1), 7-19.

[14] Liu, X. Y., Wu, J. & Zhou, Z. H. (2006), Exploratory under-sampling for class-imbalance learning, in `Proc. ICDM'.

[15] Bauer, E. & Kohavi, R. (1999), `An empirical comparison of voting classification algorithms: bagging, boosting and variants', Machine Learning 36, 105-139.

[16] Webb, G. (2000), `Multiboosting: A technique for combining boosting and wagging', Machine Learning 40(2), 159-196.

[17] Putten, P. & Someren, M. (2004), `A bias-variance analysis of a real world learning problem: the coil challenge 2000', Machine Learning 57, 177-195.

[18] Aha, D. & Kibler, D. (1991), 'Instance-based learning algorithms', Machine Learning 6, 37-66.

[19] Mitchell, T. (1997), Machine Learning, The McGraw-Hill Companies, Inc.

[20] Asuncion, A. & Newman, D. (2007), `UCI machine learning repository'.

[21] Witten, I. H. & Frank, E. F. (2005), Data Mining: Practical machine learning tools and techniques, 2 edn, Morgan Kaufmann, San Francisco.http://www.cs. waikato.ac.nz/ml/weka/.