Analyzing PETs on Imbalanced Datasets When Training and Testing Class Distributions Differ

David Cieslak and Nitesh Chawla

University of Notre Dame, Notre Dame IN 46556, USA
{dcieslak,nchawla}@cse.nd.edu

Abstract. Many machine learning applications like finance, medicine, and risk management suffer from class imbalance: cases of interest occur rarely. Further complicating these applications is that the training and testing samples might differ significantly in their respective class distributions. Sampling has been shown to be a strong solution to imbalance and additionally offers a rich parameter space from which to select classifiers. This paper is concerned with the interaction between Probability Estimation Trees (PETs) [1], sampling, and performance metrics as testing distributions fluctuate substantially. A set of comprehensive analyses is presented, which anticipate classifier performance through a set of widely varying testing distributions.

1 Introduction

Finance, medicine, and risk management form the basis for many machine learning applications. A compelling aspect of these applications is that they present several challenges to the machine learning community. The common thread among these challenges persists to be class imbalance and cost-sensitive application, which has been a focus of significant recent work [2, 3]. However, the common assumption behind most of the related works is that the testing data carries the same class distribution as the training data. This assumption becomes limiting for the classifiers learned on the imbalanced datasets, as the learning usually follows a prior sampling stage to mitigate the effect of observed imbalance. This is, effectively, guided by the premise of improving the prediction on the minority class as measured by some evaluation function. Thus, it becomes important to understand the interaction between sampling methods, classifier learning, and evaluation functions when the class distributions change.

To illustrate, a disease may occur naturally in 15% of a North American population. However, an epidemic condition may drastically increase the rate of infection to 45%, instigating differences in $P(disease)$ between the training and testing datasets. Thus, the class distribution between negative and positive classes changes significantly. Scalar evaluations of a classifier learned on the original population will not offer a reasonable expectation for performance during the epidemic. A separate, but related problem occurs when the model trained from a segment of North American population is then applied to a European population where the distribution of measured features can potentially differ significantly, even if the disease base-rate remains at the original 15%. This issue becomes critical as the learned classifiers are optimized on the sampling distributions spelled out during training to increase performance on minority or positive

T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 519–526, 2008.

class, as measured by some evaluation function. If sampling is the strong governing factor for the performance on imbalanced datasets, can we guide the sampling to have more effective generalization?

Contributions: We present a comprehensive empirical study investigating the effects of changing distributions on a combination of sampling methods and classifier learning. In addition, we also study the robustness of certain evaluation measures. We consider two popular sampling methods for countering class imbalance: undersampling and SMOTE [2,4]. To determine the optimal levels of sampling (under and/or SMOTE), we use a bruteforce wrapper method with cross-validation that optimizes on different evaluation measures like Negative Cross Entropy (NCE), Brier Score ($Brier$), and Area Under the ROC Curve ($AUROC$) on the original training distribution. The former focuses on quality of probability estimates and the latter focuses on rank-ordering. The guiding question here is – *what is more effective – improved quality of estimates or improved rank-ordering if the eventual testing distribution changes?* We use the wrapper to empirically discover the potentially best sampling amounts for the given classifier and evaluation measure. This allows us to draw observations on the suitability of popular sampling methods, in conjunction with the evaluation measures, on evolving testing distributions. We restrict our study to PETs [1] given their popularity in the literature. This also allows for a more focused analysis. Essentially, we used unpruned C4.5 decision trees [5] and considered both leaf frequency based probability estimates and Laplace smoothed estimates. We also present an analysis of the interaction between measures used for parameter discovery and evaluation. *Is a single evaluation measure more universal than the others, especially in changing distributions?*

2 Sampling Methods

Resampling is a prevalent, highly parameterizable treatment of the class imbalance problem with a large search space. Typically resampling improves positive class accuracy and rank-order [6, 7, 8, 2]. To our knowledge, there is no empirical literature detailing the effects of sampling on the quality of probability estimates; however, it is established that sampling improves rank-order. This study examines two sampling methods: random undersampling and SMOTE [9]. While seemingly primitive, randomly removing majority class examples has been shown to improve performance in class imbalance problems. Some training information is lost, but this is counterbalanced by the improvement in minority class accuracy and rank-order. SMOTE is an advanced oversampling method which generates synthetic examples at random intervals between known positive examples. [2] provides the most comprehensive survey and comparison of current sampling methods.

We search a large sampling space via wrapper [10] using a heuristic to limit the search space. This strategy first removes "excess" negative examples by undersampling from 100% to 10% in 10% steps and then synthetically adds from 0% to 1000% more positive examples in 50% increments using SMOTE. Each phase ceases when the wrapper's objective function no longer improves after three successive samplings. We use $Brier$, NCE, and $AUROC$ [11, 12] as objective functions to guide the wrapper and final evaluation metrics. Figure 1 shows the Wrapper and Evaluation framework.

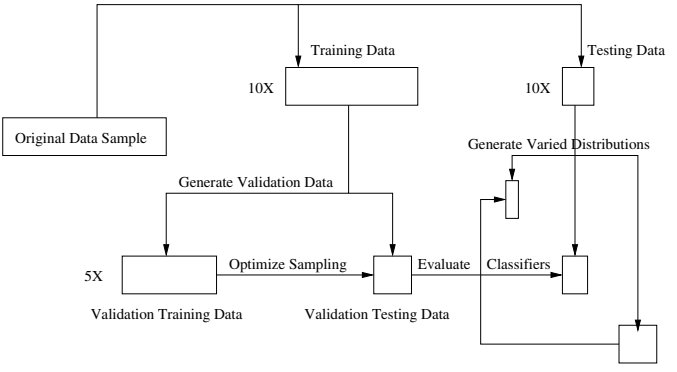

Fig. 1. Wrapper and Evaluation Framework

Table 1. Dataset Distributions. Ordered in an increasing order of class imbalance.

Dataset	Examples	Features	Class Balance
Adult [13]	48,840	14	76:24
E-State [9]	5,322	12	88:12
Pendigits [13]	10,992	16	90:10
Satimage [13]	6,435	36	90:10
Forest Cover [13]	38,500	10	93:7
Oil [14]	937	49	96:4
Compustat [10]	10,358	20	96:4
Mammography [9]	11,183	6	98:2

3 Experiments and Results

We consider performance on different samplings of the testing set to explore the range of potential distributions by exploring samplings for which $P(+) = \{0.02, 0.05, 0.1, 0.2, 0.3, ..., 0.9, 0.95, 0.98\}$. For example, suppose a given dataset has 2000 examples from class 0 and 1000 examples of class 1 in the testing set. To evaluate on $P(+) = 0.5$, 1000 class 0 examples are randomly removed from the evaluation set. We experimented on eight different datasets, summarized in Table 1.

We explore visualizations of the trends in NCE and $AUROC$ as $P(+)$ is varied. Each plot contains several different classifiers: the baseline PET [1]; sampling guided by $Brier$ (called Frequency $Brier$ Wrapper for frequency based estimates and Laplace $Brier$ Wrapper for Laplace based estimates); sampling guided by NCE; and finally sampling guided by $AUROC$ (the latter two using similar naming convention as $Brier$). In Figures 2 to 9, NCE and $AUROC$ are depicted as a function of increasing class distribution, ranging from fully negative on the left to fully positive on the right. A vertical line indicates the location of the original class distribution. $Brier$ trends are omitted as they mirror those of NCE.

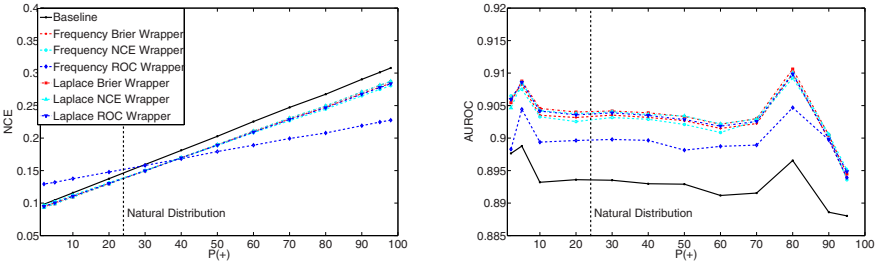

Fig. 2. Adult

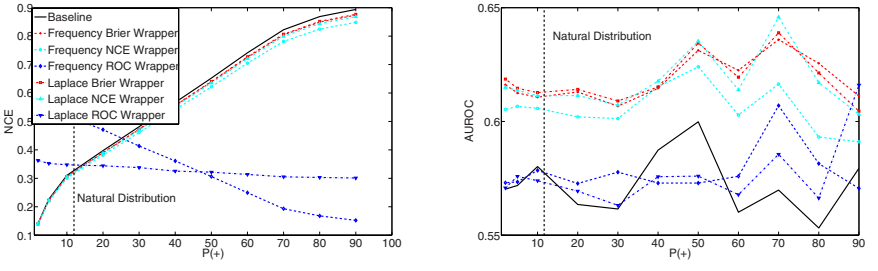

Fig. 3. E-State

Figures 2 through 9 show the experimental NCE and $AUROC$ trends as the class distribution varies. Despite the variety of datasets evaluated, some compelling general trends emerge. Throughout, we note that wrappers guided by losses generally improve NCE at and below the natural distribution of $P(+)$ as compared to $AUROC$ wrappers This implies that loss does well in optimizing NCE when the testing distribution resembles the training conditions. It is notable that in some cases, such as Figures 2, 3, 5, 6, 7, 8, & 9, that the baseline classifier actually produces better NCE scores than at least the frequency $AUROC$ wrapper, if not both $AUROC$ wrappers. The frequency $AUROC$ wrapper selected extreme levels of sampling. The reduction in NCE at low $P(+)$ indicates that using loss measures within the wrapper lowers the loss estimates for the negative class examples. That is, while the loss from the positive class may actually increase, the lower overall losses are driven by better calibrated estimates on the predominantly occurring majority class. On the other hand, classifiers learned from the $AUROC$ guided wrappers do not result in as well-calibrated estimates. $AUROC$ favors the positive class rank-order, while $Brier$ and NCE tend to treat both classes equally, which in turn selects extreme sampling levels. Thus, if NCE optimization is desired and the positive class is anticipated to occur as rarely or more rarely than in the training data, sampling should be selected according to either $Brier$ or NCE.

However, the environment producing the data may be quite dynamic, creating a shift in the class ratio and causing the minority class to become much more prevalent. In a complete paradigm shift, the former minority class might become larger than the former majority class, such as in an epidemic. Invariably, there is a cross-over point in

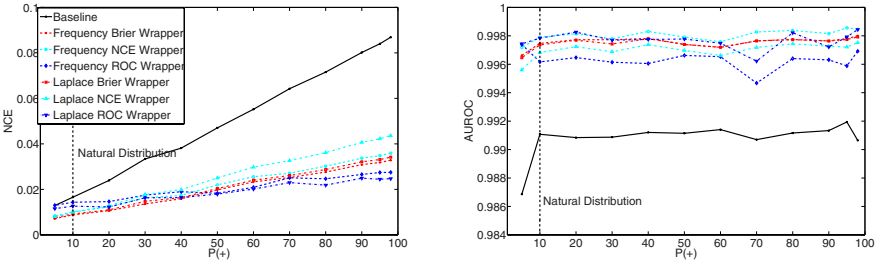

Fig. 4. Pendigits

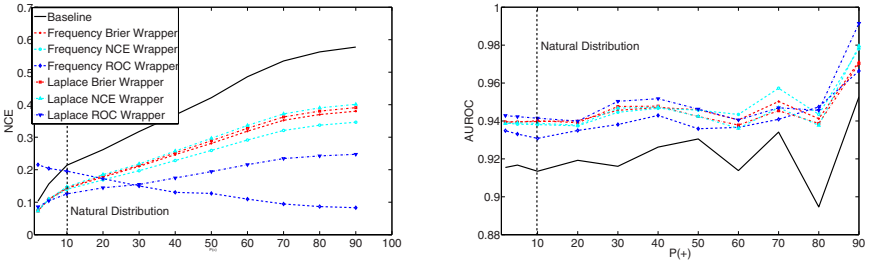

Fig. 5. Satimage

each dataset after which one of the $AUROC$ wrappers optimizes NCE values. This is logical, as $AUROC$ measures the quality of rank-order in terms of the positive class — extra emphasis is placed on correctly classifying positive examples and is reflected by the higher selected sampling levels. As the positive examples eventually form the majority of the evaluation set, classifiers producing on average higher quality positive class probability estimates will produce the best NCE. Therefore, if a practitioner anticipates an epidemic-like influx of positive examples sampling methods guided by $AUROC$ are favored.

Improvement to $AUROC$ under varied testing distributions is not as uniform. We observe that at least one loss function wrapper generally produces better $AUROC$ values in Figures 2, 3, & 4, but that an $AUROC$ wrapper is optimal in Figures 6, 7, & 9. It is difficult to declare a champion in Figures 5 & 8. It is of note that datasets with naturally larger positive classes tend to benefit (in terms of $AUROC$) from a loss wrapper, while those with naturally smaller positive classes benefit more from the $AUROC$ wrapper. As seen before, $AUROC$ guides a wrapper to higher sampling levels than $Brier$ or NCE. In the cases of relatively few positive examples (such as Forest Cover, Oil, and Mammography), a heavy emphasis during training on these few examples produces better $AUROC$ values. For the datasets with a larger set of positive examples (as in Adult, E-State, and Pendigits) from which to naturally draw, this over-emphasis does not produce as favorable a result. Therefore, in cases where there are very few positive examples, a practitioner should optimize sampling according to $AUROC$. Otherwise, $Brier$ or NCE optimization is sufficient.

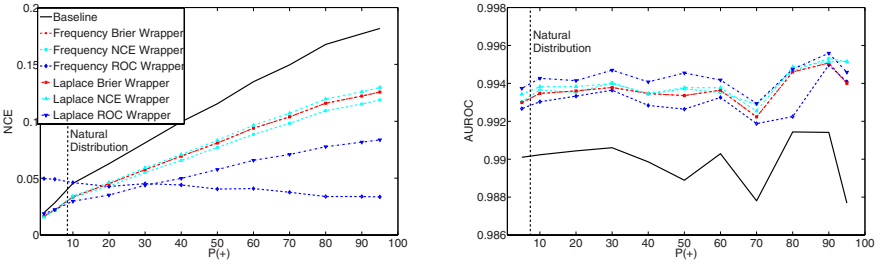

Fig. 6. Forest Cover

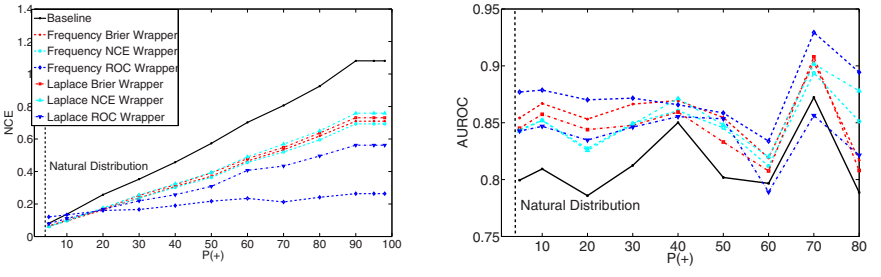

Fig. 7. Oil

The difference of characteristics between the trends in NCE and $AUROC$ is noteworthy. The NCE trends appear stable and linear. By calculating the loss on each class at the base distribution, it appears that one is able to project the NCE at any class distribution using a weighted average. $AUROC$ trends are much more violent, likely owing to the highly perturbable nature of the measure. Adding or removing a few examples can heavily impact the produced ranking. As a measure, $AUROC$ is characteristically less predictable than a loss function.

We also note that sampling mitigates the need of application of Laplace smoothing at the leaves. We can see that the baseline classifier benefits from smoothing, as also noted by other works. However, by treating the dataset for class imbalance first, we are able to counter the bias and variance in estimates arising from small leaf-sizes. The wrapper essentially searches for the ideal training distribution by undersampling and/or injecting synthetic minority class instances that lead to a reduction in loss or improvement in ranking.

Throughout Figures 2 to 9, we also note that $Brier$ and NCE loss wrappers tend to perform similarly across measures and datasets. This is not surprising as the shape of $Brier$ and NCE values are similar. We observe that the optimal sampling levels found by $Brier$ and NCE are similar, certainly more similar than to those samplings of $AUROC$. In general, NCE maintains a slight performance edge. If in the interests of time a practitioner may only experiment using one loss measure, then this study recommends using NCE, although the results found here may not apply to all domains and performance metrics.

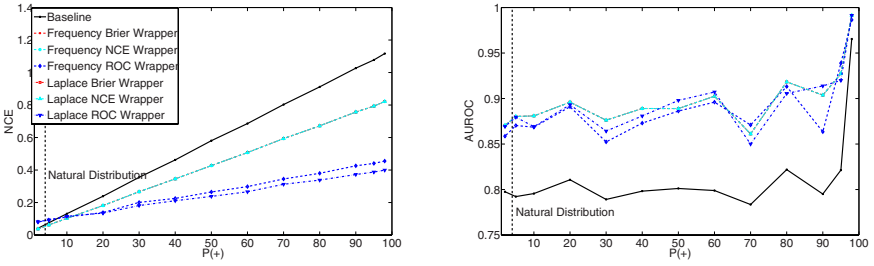

Fig. 8. Compustat

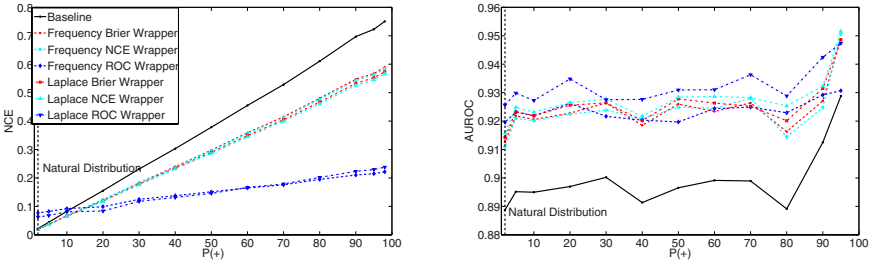

Fig. 9. Mammography

4 Conclusions

The main focus of our paper was to empirically explore and evaluate the interaction between techniques for countering class imbalance, PETs, and corresponding evaluation measures under circumstances where training and testing samples differ. In light of the questions posited in the Introduction, we make the following key observations.

- We demonstrated that it is possible to identify potentially optimal quantities of sampling by optimizing on quality of estimates or rank-order as calculated by AUROC. Almost all the wrappers demonstrated significant improvements in AUROC and reductions in losses over the baseline classifier, irrespective of the dataset. As an evaluation measure, NCE is much more stable and predictable as compared to $AUROC$. We observe NCE to change almost linearly as a function of $P(+)$, while $AUROC$ tends to change as $P(+)$ changes.
- There is a strong inter-play between undersampling and SMOTE. The wrapper determines an interaction between both the approaches by searching undersampling parameters before oversampling via SMOTE.
- It is much more difficult to anticipate the effects of a class distribution shift on $AUROC$ than it is on probability loss functions. When a dataset is highly imbalanced, we recommend guiding sampling through $AUROC$ as this places the necessary emphasis on the minority class. When class imbalance is much more moderate, NCE tends to produce an improved $AUROC$.

– While Laplace smoothing has a profound effect in improving both the quality of estimates and ranking for the baseline classifier, the advantage diminishes with sampling methods. The combination of SMOTE and undersampling improves the calibration at the leaves and thus we observed that wrapper based sampling methods are able to improve performance — lower losses and higher ranking — irrespective of smoothing at the leaves.

References

1. Provost, F., Domingos, P.: Tree Induction for Probability-Based Ranking. Machine Learning 52(3), 199–215 (2003)
2. Batista, G., Prati, R., Monard, M.: A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. SIGKDD Explorations 6(1), 20–29 (2004)
3. Chawla, N., Japkowicz, N., Kolcz, A.: Editorial: Special Issue on Learning from Imbalanced Data Sets. SIGKDD Explorations 6(1), 1–6 (2004)
4. Estabrooks, A., Jo, T., Japkowicz, N.: A Multiple Resampling Method for Learning from Imbalances Data Sets. Computational Intelligence 20(1), 18–36 (2004)
5. Quinlan, J.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1992)
6. Japkowicz, N.: The Class Imbalance Problem: Significance and Strategies. In: ICAI 2000, pp. 111–117 (2000)
7. Ling, C.X., Li, C.: Data Mining for Direct Marketing: Problems and Solutions. In: KDD, pp. 73–79 (1998)
8. Solberg, A., Solberg, R.: A Large-Scale Evaluation of Features for Automatic Detection of Oil Spills. In: ERS SAR Images IEEE Symp. Geosc. Rem., vol. 3, pp. 1484–1486 (1996)
9. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. JAIR 16, 321–357 (2002)
10. Chawla, N.V., Cieslak, D.A., Hall, L.O., Joshi, A.: Automatically countering imbalance and its empirical relationship to cost. Utility-Based Data Mining: A Special issue of the International Journal Data Mining and Knowledge Discovery (2008)
11. Buja, A., Stuetzle, W., Sheu, Y.: Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications (under submission, 2006)
12. Caruana, R., Niculescu-Mizil, A.: An Empirical Comparison of Supervised Learning Algorithms. In: ICML 2006, pp. 161–168 (2006)
13. Asuncion, A., Newman, D.: UCI machine learning repository (2007)
14. Kubat, M., Holte, R., Matwin, S.: Machine Learning for the Detection of Oil Spills in Satellite Radar Images. Machine Learning 30, 195–215 (1998)