Chapter 12
Learning from Imbalanced Data: Evaluation Matters

Troy Raeder[1], George Forman[2], and Nitesh V. Chawla[1]

[1] University of Notre Dame, Notre Dame, IN, USA
[2] HP Labs, Palo Alto, CA, USA
traeder@nd.edu, ghforman@hpl.hp.com, nchawla@nd.edu

Abstract. Datasets having a highly imbalanced class distribution present a fundamental challenge in machine learning, not only for training a classifier, but also for evaluation. There are also several different evaluation measures used in the class imbalance literature, each with its own bias. Compounded with this, there are different cross-validation strategies. However, the behavior of different evaluation measures and their relative sensitivities—not only to the classifier but also to the sample size and the chosen cross-validation method—is not well understood. Papers generally choose one evaluation measure and show the dominance of one method over another. We posit that this common methodology is myopic, especially for imbalanced data. Another fundamental issue that is not sufficiently considered is the sensitivity of classifiers both to class imbalance as well as to having only a small *number* of samples of the minority class. We consider such questions in this paper.

1 Motivation and Significance

A dataset is imbalanced if the different categories of instances are not approximately equally represented. Recent years have brought increased interest in applying machine learning techniques to difficult "real-world" problems, many of which are characterized by imbalanced data. The imbalance can be an artifact of class distribution and/or different costs of errors or examples. With an increasing influx of applications of data mining, the pervasiveness of the issues of class imbalance is becoming only more profound. These applications include telecommunications management [13], text classification [15, 22], bioinformatics [25], medical data mining [26], direct marketing [11], and detection of oil spills in satellite images [18]. These applications not only present the challenge of high degrees of class imbalance (for instance, some have less than 0.5% positives), but also the problem of small sample sizes. We assume that the positive (more interesting) class is the minority class, and the negative class is the majority class.

Let us consider a couple of cases here to underline the extreme imbalance in real-world applications. The first example is from the public Reuters RCV1

D.E. Holmes, L.C. Jain (Eds.): Data Mining: Found. & Intell. Paradigms, ISRL 23, pp. 315–331.
springerlink.com

dataset [19]. Figure 1 shows a histogram of the class distribution of 600+ classes identified in the dataset. The y-axis is the number of classes that belong to the histogram bin. The majority of classes occur less than 0.3%, and some of the classes have less than one part-per-ten-thousand in the dataset.

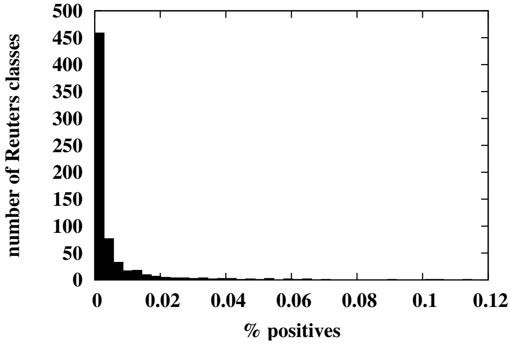


Fig. 1. Class Distribution of Reuters' Dataset

Another example is the detection of adverse drug events in a medical setting. It is extremely important to capture adverse drug events, but such events are often rare. The Institute of Medicine has encouraged incorporation of decision based tools to prevent medication errors. In our prior work, we considered prediction of such adverse drug events in Labor and Delivery [26]. The objective was to generate a classifier to identify ADE in women admitted for Labor and Delivery based on patient risk factors and comorbidities. The sample of 135,000 patients had only 0.34% instances marked as adverse drug events.

In the direct marketing domain, advertisers make money by identifying customers who will make purchases from unsolicited mailings. In this case the interesting class composes less than 1% of the population.

With a growing number of applications that are confounded by the problem of class imbalance, the question of evaluation methodology looms. We demonstrate that the choices in evaluation methodology matter substantially in order to raise the awareness to make these choices deliberately and (ideally) consistently among researchers, and to discuss frontiers of research directions.

Contribution. We address the following questions in this paper:

1. What is the effect of sample size versus class skew on the problems of learning from imbalanced data?
2. What effect does changing the class skew (making more imbalanced) have on the conclusions?
3. What is the sensitivity of validation strategies and evaluation measures to varying degrees of class imbalance?
4. Do different cross-validation strategies (10-fold or 5x2 [10]) affect the conclusions?

5. Do different evaluation measures lead us to different conclusions for the same classifiers on the same data sets?

We address the aforementioned issues by considering three different classifiers — Naive Bayes (NB), C4.5 (J48), and Support Vector Machines (SMO), and multiple datasets from a number of different domains and applications, including public data sets from UCI [3] and LIBSVM [4]. We consider both 10-fold and 5x2 cross-validation (CV) in the paper. Our evaluation methods comprise of AUC, F-measure, Precision @ Top 20, Brier score (quadratic loss on probabilistic predictions, indicative of classifier calibration), Accuracy, and the new H-measure proposed by David Hand [16]. We believe that a uniform comparison and benchmarking strategy can help innovation, achieving not only a theoretical impact but also a broad practical impact on a number of real-world domains.

2 Prior Work and Limitations

The major forefront of research in learning from imbalanced datasets has been the incorporation of sampling strategies with different learning algorithms. See the recent workshops and survey papers for a comprehensive discussion on different methods [2, 5, 23, 17, 21, 29]. Recent research has also focused on new or modified objective functions for SVMs or decision trees [8, 31, 1, 28].

We analyzed a number of papers published in the last few years on the topic of learning from class imbalance and find that researchers are very inconsistent in their choice of metric, cross-validation strategy, and benchmark datasets. Thus published studies are difficult to compare and leave fundamental questions unanswered. What is really the progress of our approach for imbalanced data? What area of the imbalanced problem space are we really addressing? What can we do as a community to ensure more real data is made available for researchers to collaborate and/or benchmark their methods on?

We give here a brief review of these recent papers on class imbalance. There is no agreement on the cross-validation strategies deployed — these range from 5x2 to 5-fold to 10-fold. Each of these can have an impact on the performance measurement, as these result in different numbers of instances in the training and testing sets. This is especially critical when there are few of the minority class instances in the dataset. The mix of performance measures is especially interesting — balanced accuracy, AUC, geometric mean, F-measure, precision, recall, and probabilistic loss measures. Particular methodologies have been shown to perform more optimally on a particular measure. The final straw man in the related work is the use of datasets. Two recent surveys on experimental comparisons of different sampling methods and classifiers (published in 2004 and 2007) have used different validation strategies (10-fold versus 5-fold), evaluation measures, and even disagreed on some of the important conclusions. Hulse et al. [17] had 8 out of 35 datasets between 1.3% and 5% of class skew. Batista et al [2] had even fewer datasets in that range, and its lowest class skew was 2.5%. Recent research on link prediction [20] has provided some insight into class skews on the order of thousands or tens-of-thousands of negative examples per positive example, but these data sets are relatively rare.

Table 1. Data sets used in this study

No.	Dataset	Examples	Features	# MinClass	% MinClass
1	Boundary (Biology)	3,505	174	140	4%
2	Breast-W (UCI)	569	30	210	37%
3	Calmodoulin (Biology)	18,916	131	945	5%
4	Compustat (Finance)	10,358	20	414	4%
5	Covtype (UCI)	38,500	54	2,747	7.1%
6	E-State (Drug Discovery)	5,322	12	636	12%
7	FourClass (LIBSVM)	862	2	307	35.6%
8	German.Numer (LIBSVM)	1,000	24	300	30%
9	Letter (UCI)	20,000	16	789	3.9%
10	Mammography (Breast Cancer)	11,183	6	223	2.3%
11	Oil (Oil Spills))	937	49	41	4%
12	Page (UCI)	5,473	10	560	10%
13	Pendigits (UCI)	10,992	16	1142	10%
14	Phoneme (Elena Project)	5,404	5	1584	29%
15	PhosS (Biology)	11,411	479	613	5%
16	Pima (UCI)	768	8	268	35%
17	Satimage (UCI)	6,435	36	625	9.7%
18	Segment (UCI)	2,310	19	330	14%
19	Splice (UCI)	1,000	60	483	48.3%
20	SVMGuide1 (LIBSVM)	3,089	4	1089	35%

While we also encounter a similar problem of limited availability of real-world datasets in this paper, we try to overcome this by artificially reducing the positive class to increase the class imbalance. We also consider a number of different real-world domains to allow for broader generalizations.

3 Experiments

We considered three different classifiers — Naive Bayes (NB), J48 (with Laplace smoothing at the leaves), and SMO using the Platt's calibration (-N 2 -M -V 2 options in WEKA). We used WEKA [27] v3.6 implementations of each to ensure repeatability. Again, our goal was not to research optimal methods of dealing with imbalance, but simply to have a set of common classifiers to illustrate the differences in evaluation methodologies and measures. Each classifier produced scores that were then plugged into a number of different measures. We used 5x2 CV and 10-fold CV. 5x2 CV performs traditional 2-fold cross-validation and repeats it with five different random splits the data; thus, each training and testing set comprises 50% of the original data. 10-fold CV splits the data into ten disjoint folds, with 90% of the data used for training (combination of 9 folds) and

10% of the data used for testing (10th fold). The folds were completely stratified, i.e. nearly the same number of positives appear in each fold; moreover, the same training and testing sets were used for each classifier to avoid any variability arising from different random seeds.

Evaluation Metrics We evaluate each classifier using a variety of measures, as indicated in the Introduction, representing the panoply appearing in recent imbalance papers. We define these measures after introducing our notation.

Assume that we are given a series of instances $x_i \in x$ and their true class labels $y_i \in y$. For two-class problems like the ones we deal with in this paper, $y_i \in \{0, 1\}$. Define the number of instances as n, the number of negative instances in the test set as n_0, and the number of positive instances as n_1. When classifying an instance x_i each classifier produces a *score* $f(x_i)$, such that instances with higher scores are deemed more likely to belong to the positive class. Many machine learning packages output scores scaled between 0 and 1, which can then be interpreted as a probability of belonging to the positive class. We assume that the cost of a misclassification error depends only on the class of the example and denote the cost of misclassifying a negative example as c_0 and the cost of misclassifying a positive example as c_1. On the basis of these scores, we define the metrics used in the paper:

- **Accuracy**: The most basic performance measure, simply the percentage of test instances that the classifier has classified correctly. For the purposes of assigning classifications to instances, we use a threshold of 0.5. That is, instances with $f(x_i) < 0.5$ are classified as negative, and all other instances are classified as positive.
- **AUC**: *AUC* quantifies the quality of the scores $f(x_i)$ in terms of rank-order. AUC is usually calculated as the empirical probability that a randomly chosen positive instance is ranked above a randomly-chosen negative instance. That is: $AUC = \frac{1}{n_0 n_1} \sum_{i|y_i=1} \sum_{j|y_j=0} I(f(x_i), f(x_j))$, where $I(x, y)$ takes on the value 1 if $f(x_i) > f(x_j)$, 1/2 if $f(x_i) = f(x_j)$ and 0 otherwise. AUC is often preferred over Accuracy for imbalanced datasets because it does not implicitly assume equal misclassification costs.
- **Brier Score**: The Brier score is the average quadratic loss on each instance in the test set: $S_{brier} = \frac{1}{n} \sum_i (f(x_i) - y_i)^2$. This quantifies the average deviation between predicted probabilities and their outcomes.
- **Precision @ Top 20**: The Precision @ Top 20 is simply the fraction of the top 20 instances (as ranked by $f(x_i)$) that are actually positive. It measures the ability of a classifier to accurately place positive instances in the most important positions, i.e. for information retrieval.
- **F-measure**: F-measure measures a classifier's effectiveness at both precision and recall. The measure we implement is known as the F_1-Measure, which is simply the harmonic mean of precision and recall. Again, we use a threshold of 0.5 to distinguish between positive and negative instances.
- **H-Measure**: H-Measure [16] is a very recently developed threshold-varying evaluation metric that is designed to overcome an inherent inconsistency in

the AUC metric. H-measure calculates the *expected loss* of the classifier (as a proportion of the maximum possible loss) under a hypothetical probability distribution $u(c)$ of the class-skew ratio $c = \frac{c_0}{c_0+c_1}$. For the purposes of this paper, we use the $beta(2,2)$ distribution suggested by Hand [16] which is given by $u(c) = 6c(1-c)$.

- **Precision-Recall Break-Even point**: A *precision-recall* (PR) curve [9] plots recall on the x-axis and precision on the y-axis as the classifier's decision threshold varies across all possible values. The precision-recall break-even point is calculated as the intersection point between the PR curve and the line `y = x`. In the event that multiple intersection points exist, the largest value is used.

The appropriateness of many of these measures has been hotly debated in the literature. Accuracy is generally regarded as a poor metric because it implicitly assumes equal misclassification costs, which is rarely true in general and never true for imbalanced problems. Additionally it requires the researcher to choose a decision threshold, often without knowledge of the domain [24]. AUC is very popular in applications involving imbalanced data, both because it does not require the choice of a decision threshold and because it is completely agnostic to class skew.

However, AUC is not without its detractors. Two of the most vocal criticisms of AUC are that it is *misleading* in cases of extreme class skew [9] and that it is an *inconsistent* measure of classification performance. We briefly address these points now, as they lead nicely into important points later in the paper. Both arguments, at their heart, deal with the relationship, or lack thereof, between AUC and actual misclassification cost.

Consider a simple test set with 9 negative examples and 1 positive example (9:1 class skew). If the examples, ranked by $f(x_i)$ have classes $\{$`0 1 0 0 0 0 0 0 0 0`$\}$, then the classifier's AUC is 0.9, the precision at the optimal decision threshold is 0.5, and the misclassification cost at the optimal threshold is c_0. A similar example can be concocted under 99:1 class skew. If ten negative examples are ranked above the single positive example, the AUC is still 0.9, but the optimal precision is 0.09, and the optimal misclassification cost is $9c_0$. Thus, two classifiers with identical AUC can incur vastly different misclassification costs, depending on the inherent difficulty of the problem under consideration. In other words, *there is no simple way to infer misclassification cost from AUC.*

Hand takes this argument one step further and shows that the actual relationship between AUC and misclassification cost is complicated and is equivalent to assuming a likelihood distribution over the possible cost ratios that depends on the classifiers being compared. Instead, he proposes to estimate misclassification cost by fixing a continuous distribution over the cost ratios and computing expected classification loss.

This is a reasonable approach except that accurate performance estimation then depends on the choice of probability distribution. In his paper, Hand proposes a Beta distribution given by $u(c) = 6c(1-c)$. There are two potential problems with this choice. First, the distribution $u(c)$ has the greatest mass near $c = 0.5$, the value that represents equal misclassification costs. Second, it is symmetric about this point, meaning that it actually assigns a likelihood of 0.5 to the possibility that the misclassification of minority class examples is *less* costly than the misclassification of majority class examples. As we will see later, this poses a problem under circumstances of extreme imbalance.

Brier score is unique among the metrics we consider in that it actually takes the magnitude of the score $f(x_i)$ into account. It seems most appropriate for situations (such as investment or betting, perhaps) where the action taken depends on the absolute confidence of the classifier in its prediction. If this information is irrelevant, and only the relative positions of the instances matter then Brier Score is an inappropriate metric, because it has a substantial impact on the rank-ordering of classifiers in our results.

3.1 Datasets

Table 1 summarizes the different datasets from different applications, and public sources such as the UCI [3] and LIBSVM [4]. Data is derived from biology [25], medicine [6, 26], finance [7], and intrusion detection. Some of these datasets were originally multi-class datasets and were converted into two class problems by keeping the smallest class in the data as minority and clumping the rest together as majority class.

The class imbalance varies from 2.3% to 48.3% (balanced). However, in our experiments we also reduced the number of minority class examples in the data, such that the class priors were artificially reduced to half of the original. That is, if the original data had 140 minority class instances, we reduced it by multiples of 5% until we had 70 (50%) minority class instances. This allowed us to consider the effect of sample size and high class skews in the experiments as well. We removed a maximum of 50% to be consistent across all the datasets; while some datasets could support further reduction, it would have severely impacted some of the datasets with few positives, such as Oil, which only has 41 examples to start with.

3.2 Empirical Analysis

We show aggregate results across all the datasets. Please note that the point here is not to compare classifiers or to state which classifier is most appropriate for a given dataset. Rather, the point is to see the sensitivity of classifiers and performance measures (and hence conclusions drawn) to different validation strategies and rates of class imbalance.

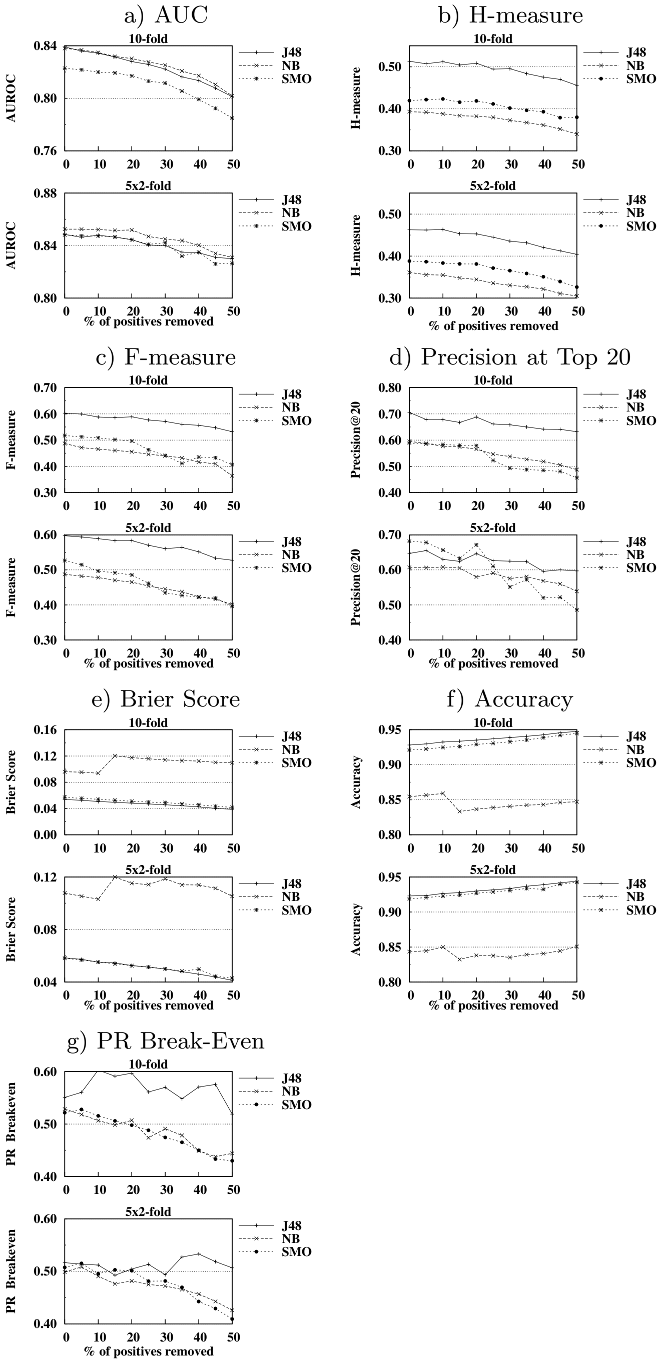
Figure 2 shows the different performance measures. The y-axis on the figure is the performance measure averaged over all datasets, and the x-axis is the increasing rate of imbalance. That is, the leftmost point (0) is the original dataset, and as we move along the x-axis, we remove x percent of the minority class. So, 10 represents removing 10% of the minority class examples prior to splitting for cross-validation.

Some interesting trends emerge from these results. Let us first consider Figure 2(a) for AUC. For each of the three classifiers, the AUC consistently drops as the imbalance increases. However, the AUC does not change nearly as much as one might expect (compare the y-axis range with the wide range of almost every other graph). This illustrates a weakness of AUC, which was pointed out by Hand [16]: the measurement of AUC depends on the relative score distributions of the positives and the negatives, which essentially depends on the classifier itself. It is independent of class priors; it is measuring only the quality of rank-order. In the absence of true costs of misclassification, AUC is relying on score distributions, which are not shifting significantly, since the feature distribution $p(x)$ for the classifier is a random subset of the original data. The change in class skew toward high imbalance is not having a significant effect. Furthermore, we see that when using 5x2, NB is the best classifier, whereas this is not observed with 10-fold. Thus, if one were to use 5x2 CV in a paper, NB may emerge as a winner, while another paper using 10-fold may discover a tie between J48 and NB. The question then is, *which one to believe?*

Figure 2(b) shows the performance with H-measure, as proposed by David Hand [16]. Hand argues the limitations of using AUC for comparing classifiers — each classifier is calibrated differently, and thus produces different score distributions. It implies that AUC is evaluating a classifier conditioned on the classifier itself, thereby resulting in different "metrics" for comparing classifiers. To that end, he proposes the H-measure, which is independent of the score distributions. It is not independent of the class priors and is sensitive to the class skew, as one would expect. This is a necessary property as the misclassification costs are related to class priors. As we shift the minority class instances to be more skewed, the class priors are changing and the evaluation measures will shift. The H-measure declines with the increasing class skew and also demonstrates a higher variance than AUC for the same classifier over different rates of class imbalance. It is also more sensitive to the size of training and testing sets, as compared to AUC.

Figure 2(c) shows the result with F-measure. Both 10-fold and 5x2 are indistinguishable in this case. The F-measure is computed by thresholding at 0.5, and then calculating the TP, FP, TN, and FN. It is simply a function of those quantities at a fixed threshold. F-measure is also very sensitive to imbalance and rapidly drops, which is not surprising as both precision and recall will deteriorate. We found F-measure exhibited a greater variance as compared to AUC.

Figure 2(d) shows Precision @ Top 20. Again the performance generally drops across imbalance. As the class imbalance increases, the expectation of a minority class example to be in the Top 20 of the probability scores (ranks) drops. Hence, the relative precision drops as the imbalance increases. There is no thresholding done, and the performance is reflective of ranking, such as one may desire in most information retrieval tasks where high recall is not essential. Furthermore, observe that with 10-fold cross-validation J48 dominated, whereas with 5x2 cross-validation the NB classifier dominated.



a) AUC

b) H-measure

c) F-measure

d) Precision at Top 20

e) Brier Score

f) Accuracy

g) PR Break-Even

Fig. 2. Performance trends at increasing levels of class imbalance

Figure 2(e) shows the result on Brier score. As a loss measure, lower loss is better. For J48 and SMO, as imbalance increases the loss decreases, which is expected given that fewer of the positive class examples are contributing to the loss function. Since there are more negative class examples, the model is calibrated better towards predicting the negative class. NB is different from the two classifiers. The blip in NB performance at 40% appears to be a random event: the high imbalance caused performance to degrade severely on the compustat dataset, which captures the rating of companies based on their financial parameters for three different years. However, the general trend of Naive Bayes corroborates the previous observations of Domingos & Pazzani [12] and Zadrozny & Elkan [30]. They have noted that Naive Bayes gives inaccurate probability estimates (but can still give good rank-ordering). Naive Bayes tends to give more extreme values, and with the shrinking minority class examples, the classifiers are becoming worse in their calibration. Since this does not affect its ability for rank-ordering, this phenomenon was not observed with AUC.

For completeness, we also included the accuracy Figure 2(f), even though it is accepted to be a weak metric for imbalanced datasets. As expected, accuracy increases with imbalance — a classifier becomes increasingly confident on the majority class. Hence, accuracy is not a useful metric for class imbalance research.

Finally, Figure 2(g) shows results for the break-even point of precision and recall. The most striking aspect of this graph is the instability of the metric under increasing imbalance for the J48 classifier. While NB and SMO generally decline in performance as the difficulty of the classification task increases, J48's performance is tremendously erratic, especially under 10-fold cross-validation. This variability serves to illustrate an important point: while performance under two-fold cross-validation may suffer from a lack of positive *training* examples, the lack of positive *test* examples in CV folds can make estimation under extreme imbalance problematic. J48 is unique in that it generally provides very coarse-grained probability estimates (based on class membership at the leaves). One result of this is that large blocks of test examples can be given the same probability estimate. As a result, small changes in classifier probability estimates can result in very large changes in the rank-ordering of positive examples. If there are few test examples, this will have a profound effect on the final performance estimate.

Summary. The results generally show that (1) greater class imbalance leads to a decay of the evaluation measure (except for accuracy), and more importantly, (2) the choice of evaluation methodology can have a substantial effect on which classifier methods are considered best. The three classifiers were ranked differently by the different evaluation measures. For example, Naive Bayes performed terribly for Brier score, and yet its rankings with respect to AUC were the best. This result underscores the importance of choosing a metric which is appropriate for the final application of the classifier. Moreover, in some cases the cross-validation strategy also has a large effect on the conclusions, especially in the case of Precision @ Top 20. With more classifiers being evaluated in a real study, the inconsistent results would multiply.

In our results, we observe several differences evaluation metrics and cross-validation methods. F-measure was more favorable to J48 versus NB or SMO. On the other hand, AUC generally found J48 and NB competitive, with a slight bias towards NB under 5x2 cross-validation. The H-measure strongly favors J48 and not so SMO and NB. Precision @ Top 20 yields a clear winner with no ties, but that winner depends on which form of cross-validation is used (J48 for 10-fold and NB for 5x2-fold). If we compare based on Brier score, NB emerged as the weakest classifier, with no clear distinction between J48 and SMO, which is not surprising given the poor calibration of NB. Finally, if we look at Precision @ Top 20, NB again was the weakest classifier, with no significant differences between SMO and J48.

These results are clear evidence that different validation methods and performance measures can result in potentially different conclusions. These variations in classifier ranking show that it is important for the community to evaluate classifiers in the light of different metrics and to be very careful when stating conclusions that may not deserve much generalization.

4 Discussion and Recommendations

We conclude with some general recommendations in the light of the results, related research, and make a call to the community for research directions, problems and questions as we strive to handle greater degrees of class imbalance.

4.1 Comparisons of Classifiers

It is evident from Figure 2 that, depending on the measure and/or the mode of validation, one can arrive at a fundamentally different conclusion about the tested classifiers. The scenario of selecting a single 'best' classifier that performs well on one chosen measure makes complete sense for more focused application settings where an optimal performance objective has been determined. But it becomes myopic or misleading for general research papers comparing methods.

Comparing the different measures sheds an interesting light. As an example, let us consider SMO at 10-fold. While it has a competitive performance in AUC, its Precision @ Top 20 suffers at a high class imbalance. This also demonstrates a potential weakness of AUC as it is looking at the entire curve. The classifier is not able to achieve a relatively higher precision in the beginning of the curve, but potentially recovers the performance along the curve, leading to a higher AUC. Now a practitioner may only be interested in the power of a classifier in ranking correct positive class predictions over the negative class, without an explicit threshold. A high AUC in this case can be misleading. A similar comparison can be drawn between Precision @ 20 versus H-measure. H-measure puts NB as the worst classifier for 5x2 but Precision @ 20 puts it as the best classifier. Such differences in the ranking of the classifiers bring out a compelling point — different classifiers have different optimal operating regions in the trade-off between the two types of errors for imbalanced data. Looking at a single metric without attention to how the classifier may be used or even the property of the data (degree of class imbalance, sample size, etc) may bring one to incorrect conclusions.

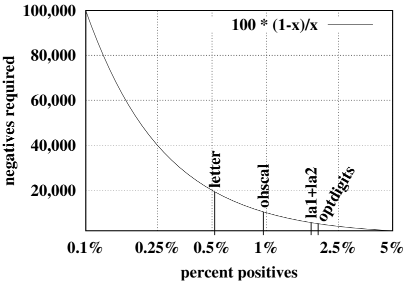


Fig. 3. The minimum number of negative cases required in a dataset in order to do research with x% positives, with a minimum of 100 positives

Effect of Sample Size. As we observed in the previous section, the limited sample size of the positive class mitigates careful experimentation and generalized conclusions. As the research community studies greater degrees of imbalance, we will need larger public benchmark datasets. How large should the datasets be? Clearly there needs to be some minimum number of positive cases in the dataset, which we discuss further in the next section. Suppose one decides that 100 positive examples are sufficient for some learning task and that they would like to perform imbalance research up to, say, 0.25% positives. Then 39,900 negative examples will be needed. Even our largest text and UCI datasets do not have anywhere near this number of negatives. Figure 3 shows the number of negatives needed for a variety of imbalance goals down to one part-per-thousand, assuming a minimum of 100 positives, which is probably a bare minimum. The figure also marks for each of our larger datasets the greatest imbalance that it can support. Keep in mind that this curve represents a lower bound. The minimum requirement on positives may need to be increased—with a proportional increase in the demand for negatives.

Sample Size and Evaluation. Consider a data set with fewer than 50 positive examples. If we do a 10-fold CV, then the number of positive training items in each fold will be no more than 45, and the testing positives will be less than or equal to 5. While this gives a reasonable (relative) sample for training, the testing set is very small, which could lead us to arrive at potentially unreliable performance estimates. The extreme scenario for 10-fold cross-validation is that there are some folds that have no positive class instances. If we do a 5x2 fold, then it would give us about 25 positive examples in training and testing. This is a much smaller size for training and will now actually effect the model calibration. By using just 50% of the dataset for training, we are indirectly preferring classifier models that can learn well from smaller samples — a perhaps unintended consequence of a methodology choice that may have little bearing for research with more balanced class distributions.

The small sample size issue is clearly confounded by the need for **internal cross-validation** to allow learning methods that perform some sort of self-calibration or parameter tuning, such as the well known Platt scaling post-processing phase for SVM, or the selection of its complexity parameter C via internal cross-validation. This internal validation becomes tricky and questionable, as the number of instances per fold are even smaller. Can the parameters then be trusted?

Null Hypothesis. If we have very few positives, not only may we be unable to determine the best method, in addition there is the possibility that we may mistake *worthless* methods for good ones. One might not think this would be a concern, but it happened in the thrombin task of the 2001 KDD Cup. The score of the winning entry achieved 0.68 AUC, and with 634 test cases, people generally believed that the test set was big enough to yield valid results. But it turned out that if each of the 117 contestants were to have submitted completely random classifiers, the *expected value* for the highest score would be slightly higher [14].

But the lesson holds especially for researchers of class imbalance. If we have small a number of positives, the possibility of getting large performance scores under the null hypothesis is remarkably high. For example, supposing we have 50 positives and 1000 negatives, the AUC critical value that must be exceeded is 0.654 in order to limit the probability to p=0.01 that our best method's score could be due only to chance—alarmingly high [15].

Test Variance. Even supposing that our methods perform well above the critical value for random classifiers, just having fewer positives in the test set leads to higher variance for most performance measurements, except accuracy or error rate. Greater variance in our test results makes it more difficult to draw research conclusions that pass traditional significance tests, such as the paired t-test or Wilcoxon rank tests.

We illustrate this point with AUC, since its known insensitivity to the testing class distribution is sometimes incorrectly taken to mean that it is acceptable to measure AUC with very few positives. We simulated a fixed classifier on various test sets, varying the number of positives and negatives. As expected, the mean AUC averaged over millions of trials was always the same, regardless of the test set. But the variance tells another story. For example, a fixed classifier that achieved mean 0.95 AUC on all test sets had the following standard deviation: 0.010 for 100:5000 positives to negatives, 0.011 for 100:500, and 0.032 for 10:500. To interpret this, the standard deviation changed little (+9%) for a shift in the class distribution from 100:5000 to 100:500, but changed a lot (+320%) when the class distribution was preserved but the number of test items was decreased in size from 100:5000 to 10:500. Furthermore, when we reduce only the positives for a large test set of 10:5000, we still get high variance (+314% of that of 100:5000). The upshot of this demonstration is that we need to have large numbers of positives in our test sets, in order to keep the variance of our test measurement low.

4.2 Towards Parts-Per-Million

Suppose ambitious researchers extend their goal to one part-per-million, 0.0001%: then 9,999,900 negatives would theoretically be needed to balance 100 positives. Such demands for labeled data are unreasonable. Not only does the effort to label by random sampling grow linearly with the total size of the dataset, it would likely also suffer from class noise that well exceeds one part-per-million. And once the price is paid to obtain all these labeled negatives, what is to be done with them? One of the most successful techniques for dealing with class imbalance is simply to discard many negatives from the training set, whether by random sampling or more involved methods. Paying a large cost to obtain a huge dataset and then throwing away a large fraction of it is somewhat nonsensical.

Thus, it appears that research under very high class imbalance cannot expect to receive randomly sampled datasets. Instead, the datasets will consist of a small number of selected positives, and a mostly unverified supply of background cases, for which the prevalence of positives is expected to be low, but non-zero. This approach has been adopted by the Information Retrieval community since the 1970's. To cope with the problem, they have developed the idea of pooled judging: all cases predicted to be positive by any of the competing methods are pooled into a union set, and then domain experts laboriously check each one to determine its ground-truth label. Once this judging is completed, one can finally score the individual methods based on their *true positives* and *false positives*, yielding F-measure, Precision, Recall, Precision@20, etc. Since no judgment is made on the majority of cases that were not retrieved by any of the methods, one cannot know the true recall, accuracy or AUC (such measures are sometimes reported by making the bold and unjustified assumption that there are no other positives).

Further complicating matters, the training positives are unlikely to be a random sample from the source distribution of positives. There is a ubiquitous assumption in machine learning that the training set is a random sample from the test distribution. But for very high class imbalance, it is unreasonable to expect a person to identify the requisite number of positives by random sampling. Instead, they will probably find positive examples via search. If they search with a single keyword query and already obtain the desired minimum number of positives, they will have appeared to meet the needs of building a labeled dataset for study. But such positives would have very low diversity. In fact, the learning problem is then simply reduced to trying to figure out the keyword query that was used (likely just a few words). Not what is intended. But such cases have occurred in practice.

4.3 Recommendations

Based on our observations and preceding discussions, we make some recommendations for the research community and practitioners who are focused on high class imbalance. While our experiments do not exhaust all possible classifiers, datasets and performance measures, they do shed light on the trends of classifiers' performances under different scenarios. We characterize the evaluation of classifiers under high imbalance as follows. Foremost, if a specific classification threshold or misclassification costs are known, then naturally one should rely on the domain expertise and target the study towards the tuned costs. Otherwise, when working in a domain confounded by high imbalance and small sample size (the parts-per-million conundrum), then we argue that of all the measures we studied, it is most appropriate to use a Precision @ top N measure. The specific operating region at very conservative decision thresholds then becomes critical.

Foh H-Measure, we can conclude that it is not heavily influenced by the performance of the classifier at the top of the operating range. Indeed, observe in Figure 2(d) that Naive Bayes has good performance with respect to Precision @ Top 20, and yet in Figure 2(b) we see that Naive Bayes receives the consistently worst H-measure scores (and well as Brier scores). Based on this, we conclude that H-measure, although perhaps appropriate for more balanced situations, is not a good candidate for judging performance at highly conservative thresholds (low false positive rate, high precision, low recall).

It is worth noting that the rank-order induced by H-measure is at all points equivalent to the rank-order induced by Accuracy. Recall that the computation of H-Measure is biased toward the notion that misclassification costs are approximately equal. At high levels of imbalance this is increasingly unlikely to be true, and this result suggests that the standard computation of H-Measure may as inappropriate as Accuracy for performance comparisons on highly imbalanced data sets. A principled study of the effect of the chosen cost-likelihood distribution on final performance estimates would make interesting future work.

One caveat with Precision @ Top N is in selecting the value of N. While in many studies a value of 20 or 100 is sufficient, it would be useful to use a common threshold so that different studies may be compared. That said, if a single N is chosen, it may be inappropriate for a study in which many more than N positives are available; the classifiers may easily fill the first N positions with only positives, and then the Precision @ Top N may not differentiate among classifiers. For instance, consider the case of 10,000 positives out of 1 million instances, which still gives a high skew of 0.01, but top 20 or top 100 may not differentiate classifiers.

5 Summary

To summarize, many of the general lessons of machine learning are amplified in research under high class imbalance. We need imbalanced datasets with very many more cases than is typically available in existing public datasets. As a community we need to converge on a validation framework with a set of evaluation metrics that is used consistently throughout. In particular, the evaluation metrics chosen need to be suitable for the problems being analyzed. To draw statistically valid conclusions and avoid overfitting, the datasets must not have too few positives, and they need to have very few labeling errors, especially in the negative/majority class. When the number of positives available is limited, the choice of 10-fold or 5x2 cross-validation can substantially affect the training sets available to the classifiers; 5x2 may penalize classifiers that have difficulty training on small samples of positives. And finally, as we move toward parts-per-million, the growing need for randomly sampled data is clearly unworkable, and the nature of the research must deal with such issues as training with positives only, leveraging a large unlabeled background dataset that may contain some positives, and perhaps Information Retrieval methods for measuring performance, where we must postpone scoring a classifier until its positive predictions have been examined by a judge.

Acknowledgements. This research was supported in part by the NSF grant EECS-0926170.

References

[1] Akbani, R., Kwek, S.S., Japkowicz, N.: Applying support vector machines to imbalanced datasets. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) ECML 2004. LNCS (LNAI), vol. 3201, pp. 39–50. Springer, Heidelberg (2004)

[2] Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations 6(1) (2004)

[3] Blake, C., Merz, C.: UCI repository of machine learning databases (1998)

[4] Chang, C., Lin, C.: Libsvm data sets, `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets`

[5] Chawla, N.V., Japkowicz, N., Kotcz, A.: Editorial: special issue on learning from imbalanced data sets. ACM SIGKDD Explorations Newsletter 6(1), 1–6 (2004)

[6] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Oversampling TEchnique. JAIR 16, 321–357 (2002)

[7] Chawla, N.V., Cieslak, D., Hall, L.O., Joshi, A.: Automatically Countering Imbalance and Its Empirical Relationship to Cost. In: DMKD (2009)

[8] Cieslak, D.A., Chawla, N.V.: Learning decision trees on unbalanced data. In: ECML (2008)

[9] Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd International Conference on Machine learning, p. 240. ACM, New York (2006)

[10] Demsar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets. JMLR 7, 1–30 (2006)

[11] Direct Marketing Association. The dmef data set library, `http://www.directworks.org/Educators/Default.aspx?id=632`

[12] Domingos, P., Pazzani, M.J.: Beyond independence: Conditions for the optimality of the simple bayesian classifier. In: ICML (1996)

[13] Ezawa, K.J., Singh, M., Norton, S.W.: Learning Goal Oriented Bayesian Networks for Risk Management. In: ICML, pp. 139–147 (1996)

[14] Forman, G.: A method for discovering the insignificance of one's best classifier and the unlearnability of a classification task. In: Data Mining Lessons Learned Workshop, ICML (2002)

[15] Forman, G., Cohen, I.: Beware the null hypothesis: Critical value tables for evaluating classifiers. In: Gama, J., Camacho, R., Brazdil, P.B., Jorge, A.M., Torgo, L. (eds.) ECML 2005. LNCS (LNAI), vol. 3720, pp. 133–145. Springer, Heidelberg (2005)

[16] Hand, D.J.: Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning 77(1), 103–123 (2009)

[17] Hulse, J.V., Khoshgoftaar, T.M., Napolitano, A.: Experimental perspectives on learning from imbalanced data. In: Ghahramani, Z. (ed.) ICML, pp. 935–942. ACM, New York (2007)

[18] Kubat, M., Holte, R., Matwin, S.: Machine Learning for the Detection of Oil Spills in Satellite Radar Images. Machine Learning 30, 195–215 (1998)

[19] Lewis, D.D., Yang, Y., Rose, T., Li, F.: RCV1: A new benchmark collection for text categorization research. Journal of Machine Learning Research 5, 361–397 (2004)

[20] Lichtenwalter, R., Lussier, J., Chawla, N.: New Perspectives and Methods in Link Prediction. In: Proceedings of KDD

[21] Mease, D., Wyner, A.J., Buja, A.: Boosted classification trees and class probability/quantile estimation. Journal of Machine Learning Research 8(3), 557–562 (2007)

[22] Mladenić, D., Grobelnik, M.: Feature Selection for Unbalanced Class Distribution and Naive Bayes. In: Proceedings of the 16th International Conference on Machine Learning, pp. 258–267 (1999)

[23] Chawla, N.V., Japkowicz, N., Kolcz, A.: Proceedings of the ICML Workshop on Learning from Imbalanced Data Sets (August 2003)

[24] Provost, F., Fawcett, T., Kohavi, R.: The case against accuracy estimation for comparing induction algorithms. In: Proceedings of the Fifteenth International Conference on Machine Learning, Citeseer, vol. 445 (1998)

[25] Radivojac, P., Chawla, N.V., Dunker, K., Obradovic, Z.: Classification and Knowledge Discovery in Protein Databases. JBI 37(4), 224–239 (2004)

[26] Tafts, L.M., et al.: Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery. JBI (2009)

[27] Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)

[28] Wu, G., Chang, E.Y.: Kba: Kernel boundary alignment considering imbalanced data distribution. IEEE TKDE 17(6), 786–795 (2005)

[29] Wu, J., Xiong, H., Wu, P., Chen, J.: Local Decomposition for Rare Class Analysis. In: Proceedings of KDD, pp. 814–823 (2007)

[30] Zadrozny, B., Elkan, C.: Learning and making decisions when costs and probabilities are both unknown. In: Proceedings KDD (2001)

[31] Zhou, Z., Liu, X.: Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE TKDE 18(1), 63–77 (2006)