Technical Report ML-TR-43, Department of Computer Science, Rutgers University
January 11, 2001

The Effect of Class Distribution on Classifier Learning

Gary M. Weiss
Rutgers University/AT&T Labs
30 Knightsbridge Rd., Piscataway, NJ 08854
gmweiss@att.com

Foster Provost
Stern School of Business, New York University
44 W. 4th St., New York, NY 10012
fprovost@stern.nyu.edu

Abstract

Many of today's large data sets must be reduced in size before invoking inductive algorithms, due to the costs associated with procuring/processing the data, and because most of these algorithms cannot handle enormous amounts of data. In these cases it is important to select the training data carefully so the impact on classifier performance is minimized. A tacit assumption behind much research on classifier induction is that the class distribution of the training data should match the "natural" distribution of the data. In this paper we analyze the relationship between training class distribution and classifier performance on 25 data sets and show that the natural distribution usually is *not* the best distribution for learning—a different class distribution should generally be chosen when the data set size must be limited. We also explain how changing the class distribution of the training set affects classifier learning and why one training distribution might be better than another.

1 Introduction

Many issues arise when classifier learning is applied to today's large-scale data sets, some of which are measured in terabytes. For example, there often are costs associated with procuring the data, storing the data, cleaning the data, and transforming the data into a form suitable for learning. For example, a common question at the start of a data mining project is: how many data records and in what proportion? Furthermore, most existing learning algorithms cannot handle huge data sets at all, and in order to run quickly the training sets must be (relatively) small. For all of these reasons, it often is necessary to limit the size of a training set. In order to minimize the impact of limiting the training-set size, it is essential that the training data be chosen carefully.

The common assumption that the naturally occurring class distribution (i.e., the relative frequency of examples of each class in the data set) is best for learning is now being questioned. This is occurring because of the increasingly common need to limit the size of large data sets and because classifiers built from data sets with highly unbalanced class distributions perform poorly on minority-class examples (as we will show). In this paper we describe and present the results from a comprehensive set of experiments designed to analyze the effect of training class distribution on classifier performance. We show that the naturally occurring class distribution usually is *not* best for training and, consequently, when the training-set size needs to be restricted, a class distribution other than the natural class distribution should be used. We also provide an explanation for how the training-set class distribution affects classifier learning and why one training distribution might be better than another.

Prior research on class distribution and classifier learning has focused on cases where large amounts of data are available but where the class distribution is highly skewed and it is very costly to misclassify minority instances. In these cases the training distribution is modified by oversampling the minority class or by undersampling the majority class, because the learning algorithm either cannot accept explicit cost information or cannot use this information effectively [Breiman, *et al.*, 1984; Drummond and Holte, 2000; Kubat and Matwin, 1997]. Our research is not restricted to highly skewed data sets and consequently one of the things we are able to show is that even when a data set's natural class distribution is nearly balanced, it still often is not best for learning. Other studies examine some aspects of the relationship between training class distribution and classifier performance, but have been limited (e.g., examine only a few data sets, do not correct for training skew when using the resultant classifier) [Catlett, 1991; Chan and Stolfo, 1998].

2 Evaluating Classifier Performance with Different Training-set Distributions

Given a fixed amount of training data, different class distributions will cause an induction algorithm to generate different classifiers. What class distribution will yield the best classifier? In order to answer this question a performance measure first must be chosen. In this research we use two performance measures. We consider only two-class problems, so the performance of a classifier can be described using the "confusion matrix" shown below. *Classification error rate* is defined as 1.0 - (TP+TN)/(TP+FP+FN+TN). Note that throughout this paper we consider the minority class to be the positive class.

	Actual Positive	**Actual Negative**
Predict Positive	True Positive (TP)	False Positive (FP)
Predict Negative	False Negative (FN)	True Negative (TN)

Error rate is the standard evaluation metric in machine-learning research. However, using this form of error rate assumes that the target class distribution is known and unchanging and, more importantly, that the error costs—the cost of a false positive and false negative—are equal. These assumptions have been criticized as being unrealistic [Provost *et al.*, 1998]. Error rate is particularly suspect as a performance measure when studying the effects of class distribution since error rate is heavily biased to favor the majority class. However, highly unbalanced problems generally have highly non-uniform error costs that favor the minority class, which is often the class of primary interest (consider medical diagnosis or fraud detection); classifiers that optimize error rate are of questionable value in these cases since they rarely will predict the minority class.

An alternative method for evaluating classifier performance is Receiver Operating Characteristic (ROC) analysis, which represents the false-positive rate on the x-axis of a graph and the true-positive rate on the y-axis. Using the terminology introduced in the confusion matrix, the true-positive rate is defined as TP/(TP+FN) and the false-positive rate as FP/(FP+TN). ROC *curves* are produced by varying the threshold on a classification model's numeric output—in our case by varying the threshold on the class-probability estimate at the leaves of a decision tree. ROC analysis and its use for machine learning are described in detail elsewhere [Swets *et al.*, 2000; Provost and Fawcett, 1998]. For our purpose, the primary advantage of ROC curves is that they illustrate the performance of a classifier without regard to class distribution or error cost.

Figure 1 shows four ROC curves, each generated from the letter-vowel data set using the same number of training examples, but with different training class distributions. As described in the previous paragraph, each point on these ROC curves corresponds to an induced decision tree plus a particular output threshold. In ROC analysis, a classifier A is better than a classifier B if it is located to the northwest of B in ROC space. The point (0,0) corresponds to the strategy of never making a positive/minority prediction and the point (1,1) to always predicting the positive/minority class.

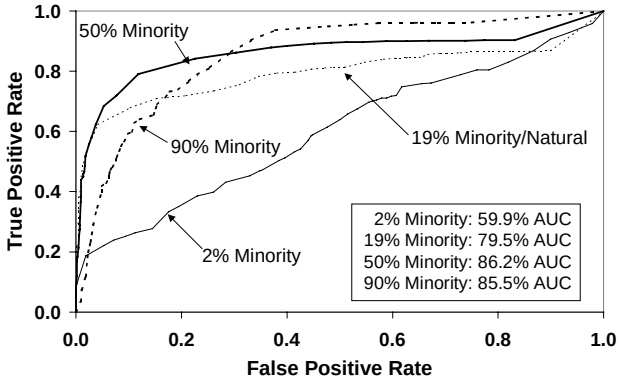

Figure 1: ROC Curves for the Letter-Vowel Data set

Observe that different training distributions perform better in different areas of ROC space. Specifically note that training with 90% minority-class examples performs substantially better than training with the natural distribution for high true-positive rates. To our knowledge such differences in performance with class distribution have never before been shown convincingly or analyzed. Unfortunately space limitations prevent us from discussing all these differences (such as why the ROC curves cross). An essential thing to note in Figure 1 is that the curve generated with the balanced training set *dominates* the curve generated with the natural distribution. This means that there is no set of target costs and class distribution for which the natural distribution is a better choice than the balanced distribution (in some cases they are almost indistinguishable).

To assess the *overall* quality of an ROC curve we use the area under the ROC curve (AUC), which is equivalent to several other statistical measures for classification and ranking [Hand, 1997]. AUC effectively averages the performance over all costs and distributions. Figure 1 includes the AUC values for the four curves. It should be kept in mind that, as shown by this figure, for *specific* cost and class distributions the best model may be not be the one that maximizes AUC. If there is not a single dominating ROC curve, multiple models can be combined so as to perform well for all costs and distributions [Provost and Fawcett, 1998]. We use AUC because we are interested in drawing conclusions across a variety of data sets as to what training distributions generally perform best.

3 Experimental Setup

We assessed the effect of class distribution on 25 data sets using both classification error rate and AUC. These data sets are described in Table 1. Of these data sets, 20 were obtained from the UCI repository [Blake and Merz, 1998] and 5 (identified with an asterisk) from researchers at AT&T. The data sets are listed in order of decreasing class imbalance, a convention we use throughout this paper. In order to simplify the presentation and analysis, data sets with more than two classes were mapped into two-class problems. This was accomplished by designating one of the original classes (generally the least frequently occurring) as the minority class and then mapping all of the remaining classes into a new class—the majority class.

#	Dataset	% Minority Examples	Dataset Size	#	Dataset	% Minority Examples	Dataset Size
1	letter-a	3.9	20000	14	network1*	29.2	3577
2	pendigits	8.3	13821	15	car	30.0	1728
3	abalone	8.7	4177	16	german	30.0	1000
4	sick-euthyroid	9.3	3163	17	breast-wisc	34.5	699
5	connect-4	9.5	11258	18	blackjack*	35.6	15000
6	optdigits	9.9	5620	19	weather*	40.1	5597
7	solar-flare	15.7	1389	20	bands	42.2	538
8	letter-vowel	19.4	20000	21	market1*	43.0	3181
9	contraceptive	22.6	1473	22	crx	44.5	690
10	adult	23.6	21281	23	kr-vs-kp	47.8	3196
11	splice-junction	24.1	3175	24	move*	49.9	3029
12	network2	27.9	3826	25	coding	50.0	20000
13	yeast	28.9	1484				

Table 1: Description of Data sets

The experiments in this paper use C4.5, a program for inducing decision trees from labeled data [Quinlan, 1993]. In order to produce ROC curves we use the Laplace correction at the leaves [Provost, *et al.*, 1998]. All experiments are run 10 times, and, except for the ROC curves in Figure 1 which were generated from a single run, all results in this paper are based on the averages over the 10 runs. Each run involves randomly selecting 25% of the minority and majority classes for testing and reserving the remaining data for training. Most experiments vary the class distribution of the training set so that the minority class accounts for between 2% and 95% of the training data.

We take two additional steps in order to ensure that the results can be compared fairly as the training distribution changes. First, the training-set size, S, for each data set is made equal to the total number of minority-class examples available for training (i.e., 75% of the total number of minority examples). Thus it is possible to generate *any* class distribution for training-set size S. Each data set selected for our study is required to contain a minimum of 200 minority-class examples in order to ensure sufficient training data.

The second step involves accounting for the differences between the training- and test-set distributions when measuring error rate—changing the training distribution will result in biased posterior class-probability estimates at the leaves of the tree and therefore may lead to inaccurate classifications (the ROC curves will not be affected). To remedy this, the probability estimate at each leaf is recomputed to take into account the difference between the training and testing distributions. The class label is then reassigned based on whether the new estimated probability of an example at the leaf belonging to the minority class is greater than or less than 50%. The revised estimates are computed as follows. Let A (B) represent the number of minority (majority) class examples at $Leaf_i$. Let *c* represent the fraction of minority-class examples in the training set divided by the fraction of minority-class examples in the test set. The revised, estimated probability of an example at $Leaf_i$ belonging to the positive/minority class is:

$$P(\text{Minority}|Leaf_i) = A / (A+cB) \quad \textbf{(1)}$$

Our results indicate that adjusting the leaf labels assigned by C4.5 yields considerable improvements in classifier error rate for training distributions other than the natural distribution (classification error rate for the natural distribution will not be affected). For example, when the training distribution is modified to contain 50% minority-class examples (and, of course, the natural distribution is used for testing), the relabeling process results in a decrease in error rate for 17 data sets, an increase for 5 data sets and no change for 3 data sets. Moreover, if we restrict our attention to those 16 data sets where the natural class ratio is greater than 3:1, then the relabeling process results in a decrease in error rate for 15 of the 16 data sets and an average decrease in error rate over the 16 data sets of 14.3%. Also, over the classifiers associated with the 16 (25) data sets, the relabeling process causes 17% (15%) of the class labels originally assigned by C4.5 to change. Prior research on the effect of class distribution on learning has not utilized this correction [Catlett, 1991; Chan and Stolfo, 1998].

One final issue concerns pruning. C4.5's pruning strategy may be adversely affected because it does not take into account the differences in class distribution between the training and test sets. Consequently, the results in this paper are based on C4.5 without pruning. Nonetheless, all experiments were repeated with pruning and exhibit similar trends.

4 Results: Natural vs. Balanced Distributions

In this section we analyze the classifiers generated from the 25 data sets listed in Table 1 using the natural and balanced class distributions. Our purpose is to show the effect of class imbalance on learning and then to contrast this to learning with no class imbalance (on the same data sets).

4.1 Learning with the Natural Class Distribution

The results of learning with the natural class distribution are shown in Table 2. The second column shows that in almost all cases more than half of all errors are the result of misclassifying minority-class examples, even though, as shown in Table 1, the minority class typically accounts for far fewer than half of the examples. The coverage column shows the average number of *training* examples that each minority-/majority-labeled leaf classifies. This column shows that leaves labeled with the minority class usually are formed from fewer training examples than those labeled with the majority class (this is not surprising since there are more majority-class examples).

	% Errs.	Coverage		Leaf ER		Example ER		Recall	
Dataset	Min.	Min.	Maj.	Min.	Maj.	Min.	Maj.	Min.	Maj.
1	65.2	6.1	12.1	7.2	0.5	12.6	0.3	87.4	99.7
2	32.2	28.6	333.9	22.0	1.1	11.8	2.2	88.2	97.8
3	72.0	4.0	49.1	56.8	7.1	77.4	2.9	22.6	97.1
4	49.2	8.1	53.9	16.2	1.6	16.0	1.7	84.0	98.3
5	51.0	2.0	6.4	48.1	5.2	49.1	5.0	50.9	95.0
6	60.8	6.8	7.0	7.0	1.2	11.5	0.7	88.6	99.3
7	69.4	2.2	6.8	66.2	13.9	81.7	6.8	18.3	93.2
8	59.3	3.8	1.4	15.9	5.1	21.7	3.6	78.3	96.4
9	51.1	2.2	4.0	64.1	18.8	64.9	18.2	35.1	81.8
10	59.1	3.4	1.9	33.0	12.3	41.5	8.9	58.5	91.1
11	55.1	7.0	12.2	12.9	4.8	15.3	4.0	84.7	96.0
12	60.0	4.7	12.7	44.4	19.7	54.5	14.0	45.5	86.0
13	70.7	11.8	35.6	37.1	21.0	59.0	10.0	41.0	90.0
14	59.6	4.7	12.0	41.2	19.7	50.8	14.3	49.2	85.7
15	70.4	4.4	13.1	5.7	5.2	12.6	2.3	87.4	97.7
16	55.2	2.1	1.8	55.9	24.9	60.9	21.3	39.1	78.7
17	53.7	11.8	29.0	8.0	4.9	9.5	4.2	90.5	95.8
18	80.2	159.1	359.0	29.8	27.7	63.3	8.7	36.7	91.3
19	48.7	5.1	7.2	40.8	26.9	39.6	27.9	60.5	72.1
20	86.3	1.6	0.5	22.5	33.8	65.0	7.5	35.0	92.5
21	55.4	6.0	3.3	26.6	22.4	31.0	19.0	69.0	81.0
22	55.3	3.6	2.2	21.4	19.5	25.5	16.2	74.5	83.8
23	67.4	40.5	78.7	0.5	0.6	0.7	0.5	99.3	99.5
24	62.9	2.7	0.6	21.2	27.6	31.3	18.4	68.7	81.6
25	64.1	2.7	1.7	23.9	31.0	35.9	20.1	64.1	79.9
Ave.	60.6	13.4	41.8	29.1	14.3	37.7	9.5	62.3	90.5
Median	59.6	4.7	7.2	23.9	13.9	35.9	7.5	64.1	92.5

Table 2: Results of Learning with the Natural Distribution

Table 2 also provides detailed, class-specific error-rate (ER) statistics for the classifiers. The "Leaf ER" column specifies the error rates for the *leaves* that predict the minority and majority classes and the "Example ER" column the error rates for the *test examples* belonging to the minority and majority classes. The "Recall" column shows the percentage of the total minority- and majority-class examples that are correctly classified. This same information is presented graphically in Figure 2 using a scatter plot.

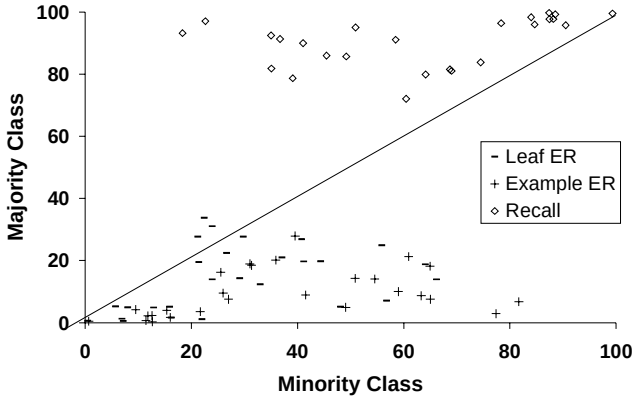

Figure 2: Results using Natural Class Distribution

Examination of the error-rate results shows that classifiers perform much worse on the minority class than on the majority class. Further, note that on average the classifiers correctly classify 62.3% of the minority examples but 90.5% of the majority. The results can be summarized as follows:

although the classifier *leaves* generally are less accurate at predicting the minority class (based on Leaf ER), the classifier performs even worse at classifying the minority *examples* (based on Example ER). This is possible because the leaves predict the minority class far less often (cf. Recall).

4.2 Learning with a Balanced Class Distribution

For comparison we next removed the class imbalance from all data sets, by including equal numbers of minority- and majority-class examples in the training and test sets. (The "minority" class refers to the class that occurs less frequently in the *natural* distribution.) The results are depicted in Figure 3.

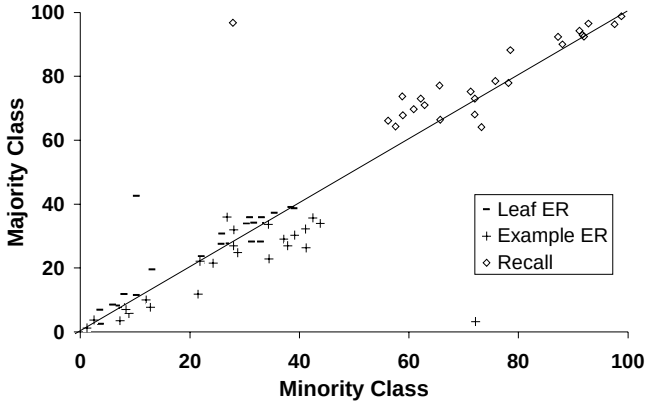

Figure 3: Results using Balanced Class Distribution

Comparing the results from Figure 3 with those from Figure 2 shows that the minority and majority values for all three measures become more similar when the class imbalance is removed. However, even without any training and test imbalance there is a striking pattern when comparing the minority- and majority-class results. Namely, for 21 of 25 data sets the minority-class examples have the higher error rate, for 19 of 25 data sets the leaves predicting the minority class have the lower error rate, and for 21 of 25 data sets the minority class has the higher recall. In summary, *when learning from a balanced class distribution the classifiers generally come up with fewer but more accurate classification rules for the minority class than for the majority class.*

Although it certainly deserves further analysis, we believe this surprising difference exists because the minority class often comprises a more homogeneous set of entities, while the majority class often corresponds to "everything else." For example, in fraud detection the minority class corresponds to illicit activities while the majority class corresponds to all other activities.

5 Results: What Training Distribution is Best?

Next we varied the training-set distributions for all 25 data sets, so that the minority class accounted for the following percentages of the data: 2%, 5%, 10%, 20%, … 80%, 90%, and 95%. Due to space limitations only the results for some of these training distributions are shown in Table 3. For each distribution the average error rate and AUC value over the 10 runs are shown. The optimal values (minimum for error rate and maximum for AUC) are underlined and displayed in bold; the vertical bars indicate the relative position of the natural distribution. By comparing the position of the optimal values with the vertical bars one can see the relationships between the optimal and natural training-set distributions.

	Nat	Error Rate using Specified Training Distribution (expressed as % minority)									AUC when Training using Specified Distribution (expressed as % minority)									Relative Improv. (%)	
Dataset	Distr.	Nat	2	5	10	20	30	40	50	60	Nat	2	5	10	30	50	70	90	95	ER	AUC
letter-a	3.9	2.8	2.9	**2.5**	2.9	3.1	3.6	5.3	5.3	6.7	79.3	74.5	81.9	86.4	92.7	94.2	94.8	**95.1**	93.2	11.4	19.9
pendigits	8.3	3.7	5.8	4.0	3.7	**3.5**	3.6	3.7	4.1	4.2	96.3	90.6	95.7	97.2	**97.8**	97.7	**97.8**	97.2	95.7	3.0	1.6
abalone	8.7	10.7	**9.0**	9.2	11.6	12.8	15.9	19.6	20.7	22.3	69.4	59.5	60.5	71.6	75.9	**76.4**	73.8	71.2	67.0	16.0	10.1
sick-euthyroid	9.3	4.5	6.9	5.4	**3.8**	4.4	7.2	6.4	9.1	11.0	92.6	80.1	91.8	93.8	94.2	**95.6**	95.3	92.9	93.3	14.3	3.2
connect-4	9.5	10.7	**7.7**	8.8	10.9	15.0	19.1	23.5	27.4	31.8	72.3	66.2	68.4	73.8	76.9	78.8	**79.3**	76.5	74.3	28.1	9.7
optdigits	9.9	4.9	9.1	7.2	5.6	3.1	**2.5**	2.9	3.3	3.7	77.8	60.0	68.6	76.5	92.6	95.7	96.5	96.4	**97.1**	49.6	24.0
solar-flare	15.7	19.2	**16.2**	17.5	19.5	21.4	21.6	27.2	27.6	29.9	**66.3**	61.1	63.2	61.7	64.5	62.9	64.1	64.6	63.9	15.4	0.0
letter-vowel	19.4	11.8	16.5	14.2	12.8	**11.7**	12.2	12.8	14.4	15.7	79.7	62.0	68.5	74.1	82.3	85.3	**86.6**	85.6	84.0	0.3	8.7
contraceptive	22.6	31.7	**23.9**	26.1	25.8	28.7	31.3	34.9	39.3	42.3	59.0	54.5	59.1	61.9	64.0	63.4	65.2	**65.3**	61.2	24.8	10.7
adult	23.6	18.2	19.2	18.3	**17.5**	17.8	18.8	19.7	20.6	22.8	82.5	80.2	79.6	80.9	82.0	83.6	**83.9**	**83.9**	82.1	3.6	1.7
splice-junction	24.1	8.3	20.5	14.2	11.5	9.0	9.2	8.3	**8.2**	10.6	91.0	75.7	83.4	85.5	91.6	93.8	94.0	**95.5**	93.2	1.0	4.9
network2	27.9	27.1	27.6	26.4	26.3	**25.9**	27.1	27.8	30.0	33.4	70.7	62.7	69.3	69.0	70.6	70.5	**70.9**	69.2	69.2	4.6	0.3
yeast	28.9	27.0	28.7	28.7	27.3	27.1	**25.9**	27.1	28.3	30.0	69.8	52.2	52.5	63.0	70.9	71.6	**71.9**	69.3	60.1	3.9	3.0
network1	29.2	27.6	27.1	27.3	26.9	**26.7**	27.8	28.9	29.9	32.7	**70.9**	66.6	67.9	69.9	69.2	70.1	68.8	68.5	68.8	3.3	0.0
car	30.0	9.5	23.0	19.3	15.4	11.2	8.5	8.0	**7.6**	7.7	87.1	72.1	76.5	78.4	89.7	92.5	91.9	**93.1**	90.1	19.8	6.9
german	30.0	33.8	29.8	29.8	31.7	**30.1**	33.3	36.2	35.8	41.0	62.9	57.4	61.0	63.0	64.5	**67.0**	65.3	63.9	65.2	10.9	6.5
breast-wisc	34.5	7.4	19.1	14.4	9.1	7.4	7.4	**6.7**	6.8	7.0	96.2	88.4	92.7	94.3	96.2	96.4	**97.5**	95.1	95.0	10.0	1.4
blackjack	35.6	**28.1**	31.1	30.4	29.9	29.1	28.2	28.2	28.2	29.0	70.2	58.2	60.2	62.8	69.6	**72.1**	70.0	59.4	54.5	0.0	2.7
weather	40.1	**33.0**	38.8	36.8	35.3	33.3	33.3	34.5	34.6	36.1	**74.3**	67.8	72.2	72.0	73.5	73.5	73.7	72.2	70.5	0.0	0.0
bands	42.2	32.3	36.3	36.2	36.2	32.3	33.8	**31.0**	34.2	32.7	**61.9**	60.6	55.8	54.8	59.8	59.9	60.3	51.3	52.5	4.0	0.0
market1	43.0	26.7	35.9	32.8	29.5	27.0	**25.8**	26.1	26.4	28.0	80.9	71.7	76.1	79.4	80.7	**81.2**	81.1	80.8	77.3	3.3	0.4
crx	44.5	21.0	36.0	31.4	29.1	22.4	21.6	20.5	21.2	**19.4**	85.4	77.1	76.4	77.2	83.4	83.3	84.3	84.3	**86.3**	7.8	1.1
kr-vs-kp	47.8	1.3	11.5	6.4	3.5	1.9	1.7	**1.1**	1.2	1.3	99.8	94.6	97.7	99.0	99.8	**99.9**	99.7	98.9	98.8	12.0	0.1
move	49.9	**27.5**	45.2	42.3	37.6	32.9	30.4	29.4	28.6	29.3	**74.8**	56.1	61.8	65.2	70.5	74.0	74.4	71.0	67.7	0.0	0.0
coding	50.0	33.5	46.0	43.2	39.8	36.4	34.5	33.5	**33.1**	33.5	68.5	61.7	61.9	63.5	67.1	69.0	**69.4**	67.6	65.5	1.2	1.3
Ave.	27.5	18.5	22.9	21.3	20.1	19.0	19.4	20.1	21.0	22.5	77.6	68.5	72.1	75.0	79.2	80.3	80.4	78.8	77.1	7.3	4.7
Median	28.4	18.8	23.0	20.3	19.8	20.2	20.5	22.0	23.8	25.4	74.8	66.2	68.6	73.8	76.9	78.8	79.3	76.5	74.3	4.8	1.7

Table 3: Optimal Training Distributions for Error rate and AUC

The results in Table 3 show that, for both error rate and AUC, training with the natural distribution seldom is optimal.† Furthermore, as indicated by the last two columns, training using the optimal distribution instead of the natural distribution leads to a substantial improvement for most of the data sets with large class imbalances. Inspection of the error-rate results shows that the optimal distribution does not differ from the natural distribution in any consistent way—sometimes it includes more minority examples and sometimes fewer. There is some correlation between the optimal and natural distributions, but far less than one might expect.

The results for AUC show that the optimal distribution is shifted to include more minority-class examples than the natural distribution. Since, unlike error rate, AUC is not affected by the test set class distribution, and because it averages over *all* distributions, this shift is not surprising—one might expect the optimal distribution to be near the 50-50 distribution. The results indicate the shift goes past the 50-50 point, something we discuss in the next section.

6 Discussion

In this section we explain the main results from Sections 4 and 5 and try to provide a better understanding of the relationship between class distribution and classifier learning.

6.1 Why is the Minority-class Error Rate Higher?

The results from Section 4.1 clearly demonstrate that classifiers tend to have a higher error rate on the minority class than on the majority class. We provide several component reasons for this behavior. The first reason, although simple, is subtle and often overlooked: *there are many more majority than minority-class test instances.* Thus, false-positive predictions are (implicitly) given more weight in error-rate-based assessments. For example, imagine a randomly generated and labeled decision tree that is evaluated on a test set with two classes, where the class ratio is 9:1. In this situation the leaves predicting the minority class will have an expected error rate of 90% while the majority-class leaves will have an expected error rate of only 10%!

The second reason that the minority class has a higher error rate is that the class "priors" in the natural training distribution are biased strongly in favor of the majority class. Thus, with equivocal evidence many classifiers will predict the majority class, which leads to higher error rates on minority-class examples.

The third reason is also related to the minority class having fewer training examples than the majority class. The coverage statistics in Table 2 show that a consequence of having fewer minority-class examples is that each rule that predicts the minority class is formed from many fewer training examples, on average, than the rules that predict the majority class. *Small disjuncts* are those rules that cover few training examples and have been shown to have a higher error rate than large disjuncts [Holte, *et al.*, 1989; Weiss and Hirsh, 2000]. Thus, the minority-class rules have a higher error rate because they suffer from this "problem of small disjuncts."

We can gain a better understanding of the problem of small disjuncts by relating it to the class distribution. Surprisingly, the reason why small disjuncts have a higher error rate than large disjuncts is due only partly to the fact that small disjuncts are based on a smaller number of training examples (i.e., the third reason). As it turns out, the first and second reasons described in this section are also factors. That is, part of the reason small disjuncts have a higher error rate than large disjuncts is because they over-represent the minority class—which tends to have a higher error rate than the majority class. We quantify the impact of each of these factors in the expanded version of this paper.

6.2 Why Isn't the Natural Class Distribution Best?

In Section 5 we saw that the natural distribution is often not best for learning, for either error rate or AUC. We are now able to provide some insight into why one training class distribution is better than another and why the natural distribution is often not best (although a thorough explanation requires further study).

We begin by looking at the learning curves for the minority and majority classes. The learning curves for the contraceptive and optdigits data sets are shown below in Figure 4. For these data sets, as for all 25 data sets, the learning curve for the minority class is on top, the overall curve in the middle and the majority-class learning curve on bottom. Thus, the test examples belonging to the minority class always have a higher error rate than those belonging to the majority class. Because the training set sizes for experiments that vary the class distributions are set to equal ¾ of the number of minority class examples (as described in Section 3), the training set sizes for the optdigits and contraceptive datasets correspond to a value of x=7% and 17%, respectively, in Figure 4. Note that at around these values the minority-class learning curve for the optdigits data set shows dramatic improvement, while for the contraceptive data set it shows only a slight improvement..

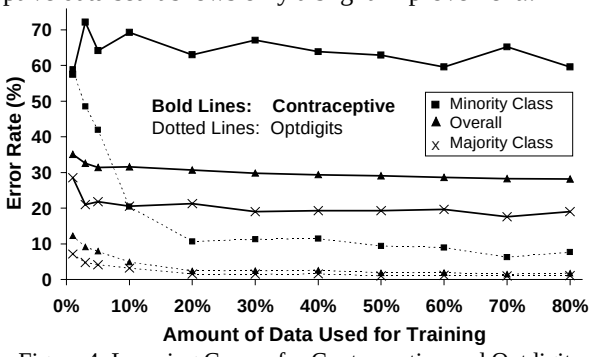

Figure 4: Learning Curves for Contraceptive and Optdigits

The dramatic improvement in optdigits' minority-class learning curve explains why, for error rate, its optimal training distribution (30% minority) includes far *more* minority examples than its natural distribution (4.9% minority); the slight improvement in the contraceptive data set's minority-class learning curve explains why its optimal distribution (2% minority) includes far *fewer* minority examples than its natural distribution (31.7% minority).

† For readers concerned with the multiple-comparisons problem, note that in almost every case there is a clear trend across the rows, with a single minimum for error rate and maximum for AUC.

More specifically, for most of the data sets we studied, the minority-class learning curves begin with a much higher error rate than the majority-class learning curves; they show more rapid improvement, but still plateau at a later point. One practical consequence of this behavior is that once the majority class has enough training data so its learning curve flattens, it makes sense to then begin adding only minority-class examples, until its learning curve also flattens.

Note that for AUC the optimal distribution for the contraceptive data set includes more minority-class examples than the natural distribution (in contrast to the error-rate case). This is because with AUC the majority class does not have greater weight than the minority class, and while the improvement in the minority-class learning curve may appear slight, at times it shows more improvement than the improvement in the majority-class learning curve.

We have just provided a qualitative analysis of why one training distribution is better than another, but a quantitative analysis is also possible. Imagine that the training-set size for the optdigits data set corresponds to the value at x=10% in Figure 4 and one example can be added to the training set. If the ratio of majority- to minority-class test examples is *X*:1 and error rate is the performance measure, then a minority-class example should be added only if the slope (i.e., improvement) of the minority-class curve at this point is at least *X* times that of the majority-class curve. For other performance metrics a similar analysis is possible.

7 Conclusion and Future Work

In this paper we showed that, and explained why, classifiers perform much worse on minority-class examples than majority-class examples. We also showed that by modifying the training-set class distribution one can usually improve the overall performance of the classifier—sometimes dramatically. We showed this for two quite different classifier performance measures, error rate and AUC. Thus, in cases where the training-set size must be limited, one can build better-performing classifiers by using a distribution other than the natural distribution. In practice, we suggest that a progressive, adaptive, sampling strategy be developed that incrementally requests new examples based on the improvement in classifier performance due to the recently added minority- and majority-class examples. This information can be estimated by using cross-validation.

Another progressive sampling strategy suggested by our analysis involves selecting training data based on the current error rate for each rule, so that more data is provided to the rules/disjuncts with higher error rates. As shown in this paper, this would typically provide more data for the minority-class examples and small disjuncts. Thus, this strategy would combat the problem with small disjuncts and move the training set toward a more balanced class distribution. It would be interesting to see how much of the effect of techniques such as boosting and active sampling can be explained by their modifying the training class distribution.

Due to space limitations we have presented only our main results. We have observed similar results with pruning, when using a third classifier performance metric, and when using a more sophisticated leaf probability estimate that factors in the class imbalance.

We hope these results help researchers and practitioners to understand better the relationship between the training class distribution and classifier performance, and to learn more effectively from large data sets in situations where the training-set size must be limited.

Acknowledgments

We would to thank Brian Davison, Chris Mesterharm and Matthew Stone for their comments on this paper and Haym Hirsh for the comments and feedback he provided throughout this research.

References

[Breiman *et al.*, 1984] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and Regression Trees. Belmont, CA: Wadsworth International Group.

[Blake and Merz, 1998] Catherine Blake and Christopher Merz. UCI repository of machine learning databases, [http://www.ics.uci.edu/~mlearn/MLRepository.html], University of California, Dept. of Computer Science.

[Catlett, 1991]. Jason Catlett. Megainduction: machine learning on very large databases. Ph.D. Thesis. University of Technology, School of Computer Science, Sydney, Australia, 1991.

[Chan and Stolfo, 1998] Phillip Chan and Salvatore Stolfo. Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 164-168, New York, NY, August 1998. AAAI Press.

[Dummond and Holte, 2000] Chris Drummond and Robert C. Holte. Exploiting the cost (in)sensitivity of decision tree splitting criteria. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 239-246.

[Hand, 1997] David J. Hand. *Construction and Assessment of Classification Rules*. John Wiley and Sons, 1997.

[Holte *et al.*, 1989] Robert C. Holte, C., L.E. Acker, and B. W. Porter. Concept learning and the problem of small disjuncts. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 813-818. San Mateo, CA: Morgan Kaufmann.

[Kubat and Matwin, 1997] Miroslav Kubat and Stan Matwin. Addressing the curse of imbalanced training sets: one-sided selection. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179-186.

[Provost *et al.*, 1998] Foster Provost, Tom Fawcett, and Ron Kohavi. The case against accuracy estimation for comparing classifiers. In *Proceedings of the Fifteenth International Conference on Machine Learning*.

[Provost and Fawcett, 1998] Foster Provost and Tom Fawcett. Robust classification systems for imprecise environments. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 706-713.

[Quinlan, 1993] J. R. Quinlan. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

[Swets *et al.*, 2000] John Swets, Robyn Dawes, and John Monahan. Better decisions through science. Scientific American, October 2000: 82-87.

[Weiss and Hirsh, 2000] Gary M. Weiss and Haym Hirsh. A quantitative study of small disjuncts, *In Proceedings of the Seventeenth National Conference on Artificial Intelligence*, 665-670. Menlo Park, CA: AAAI Press.