Toward Scalable Learning with Non-uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection

Philip K. Chan
Computer Science
Florida Institute of Technology
Melbourne, FL 32901
pkc@cs.fit.edu

Salvatore J. Stolfo
Department of Computer Science
Columbia University
New York, NY 10027
sal@cs.columbia.edu

Abstract

Very large databases with skewed class distributions and non-uniform cost per error are not uncommon in real-world data mining tasks. We devised a multi-classifier meta-learning approach to address these three issues. Our empirical results from a credit card fraud detection task indicate that the approach can significantly reduce loss due to illegitimate transactions.

Introduction

Very large databases with skewed class distributions and non-uniform cost per error are not uncommon in real-world data mining tasks. One such task is credit card fraud detection: the number of fraudulent transactions is small compared to legitimate ones, the amount of financial loss for each fraudulent transaction depends on the amount of transaction and other factors, and millions of transactions occur each day. A similar task is cellular phone fraud detection (Fawcett & Provost 1997). Each of these three issues has not been widely studied in the machine learning research community.

Fawcett (1996) summarized the responses to his inquiry on learning with skewed class distributions. The number of responses was few given skewed distributions are not rare in practice. Kubat and Matwin (1997) acknowledged the performance degradation effects of skewed class distributions and studied techniques for removing unnecessary instances from the majority class. Instances that are in the borderline region, noisy, or redundant are candidates for removal. Cardie and Howie (1997) stated that skewed class distributions are "the norm for learning problems in natural language processing (NLP)." In a case-based learning framework, they studied techniques to extract relevant features from previously built decision trees and customize local feature weights for each case retrieval.

Error rate is commonly used in evaluating learning algorithms; cost-sensitive learning has not been widely investigated. Assuming the errors can be grouped into a few types and each type incurs the same cost, some studies (e.g., (Pazzani *et al.* 1994)) proposed algorithms that aim to reduce the total cost. Another line of cost-sensitive work tries to reduce the cost in using a classifier. For instance, some sensing devices are costlier in the robotics domain (Tan 1993). Fawcett and Provost (1997) considered non-uniform cost per error in their cellular phone fraud detection task and exhaustively searched (with a fixed increment) for the Linear Threshold Unit's threshold that minimizes the total cost.

Until recently, researchers in machine learning have been focusing on small data sets. Efficiently learning from large amounts of data has been gaining attention due to the fast growing field of data mining, where data are abundant. Sampling (e.g., (Catlett 1991)) and parallelism (e.g., (Provost & Aronis 1996)) are the two main directions in scalable learning. Much of the parallelism work focuses on parallelizing a particular algorithm on a specific parallel architecture. That is, a new algorithm or architecture requires substantial amount of parallel programming work.

We devised a multi-classifier meta-learning approach to address these three issues. Our approach is based on creating data subsets with the appropriate class distribution, applying learning algorithms to the subsets independently and in parallel, and integrating to optimize cost performance of the classifiers by learning (meta-learning (Chan & Stolfo 1993)) from their classification behavior. That is, our method utilizes all available training examples and does not change the underlying learning algorithms. It also handles non-uniform cost per error and is cost-sensitive during the learning process. Although our architecture and algorithm-independent approach is not as efficient as the fine-grained parallelization approaches, it allows different "off-the-shelf" learning programs to be "plugged" into a parallel and distributed environment with relative ease. Our empirical results for the credit card fraud problem indicate that our approach can significantly reduce loss due to illegitimate transactions.

This paper is organized as follows. We first describe the credit card fraud detection task. Next we examine the effects of training class distributions on the performance. We then discuss our multi-classifier meta-learning approach and empirical results. In closing, we

summarize our results and directions.

Credit Card Fraud Detection

When banks lose money because of credit card fraud, card holders partially (possibly entirely) pay for the loss through higher interest rates, higher membership fees, and reduced benefits. Hence, it is both the banks' and card holders' interest to reduce illegitimate use of credit cards, particularly when plastic is prevalent in today's increasingly electronic society. Chase Manhattan Bank provided us with a data set that contains half a million transactions from 10/95 to 9/96, about 20% of which are fraudulent (the real distribution is much more skewed (*fortunately*)—the 20:80 distribution is what we were given after the bank's filtering).

Due to the different dollar amount of each credit card transaction and other factors, the cost of failing to detect different fraudulent transactions is not the same. Hence we define:

$$AverageAggregateCost = \frac{1}{n}\sum_{i}^{n} Cost(i)$$

where $Cost(i)$ is the cost associated with transaction i and n is the total number of transactions. After consulting with a bank representative, we settled on a simplified cost model (the cost model used by the bank is more complex and is still evolving). Since it takes time and personnel to investigate a potential fraudulent transaction, an *overhead* is incurred for each investigation. That is, if the amount of a transaction is smaller than the overhead, it is not worthwhile to investigate the transaction even if it is suspicious. For example, if it takes ten dollars to investigate a potential loss of one dollar, it is more economical not to investigate. Therefore, assuming a fixed *overhead* and considering:

[Number of instances]	Actual Positive (fraudulent)	Actual Negative (legitimate)
Predicted Positive (fraudulent)	True Positive (*Hit*) [a]	False Positive (*False Alarm*) [b]
Predicted Negative (legitimate)	False Negative (*Miss*) [c]	True Negative (*Normal*) [d]

we devised the following cost model for each transaction:

$Cost(FN)$	*tranamt*
$Cost(FP)$	*overhead* if *tranamt* > *overhead* or 0 if *tranamt* ≤ *overhead*
$Cost(TP)$	*overhead* if *tranamt* > *overhead* or *tranamt* if *tranamt* ≤ *overhead*
$Cost(TN)$	0

where *tranamt* is the amount of a credit card transaction. Furthermore, we define the false-negative rate as $\frac{c}{a+c}$ and the false-positive rate as $\frac{b}{b+d}$. Based on this cost model, we next study the effects of training class distributions on performance.

Effects of Training Distributions

Experiments were performed to study the effects of training class distributions on the credit card cost model. We use data from the first 10 months (10/95 - 7/96) for training and the 12th month (9/96) for testing. In order to vary the fraud distribution from 10% to 90% for each month, we limit the size of the training sets to 6,400 transactions, which are sampled randomly without replacement. Four learning algorithms (C4.5 (Quinlan 1993), CART (Breiman *et al.* 1984), RIPPER (Cohen 1995), and BAYES (Clark & Niblett 1989)) were used in our experiments.

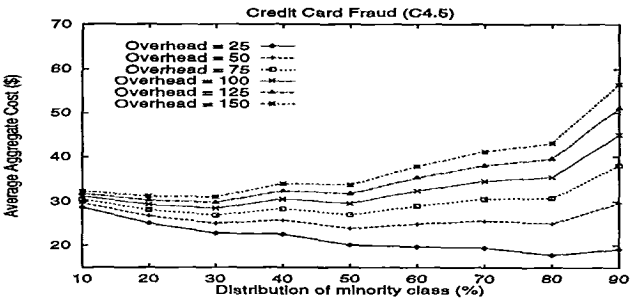

Figure 1: Training distribution vs. the credit card fraud cost model

The results are plotted in Figure 1 (due to space limitations, only results from C4.5 are shown—the other algorithms behave similarly). Each data point is an average of 10 classifiers, each of which is generated from a separate month. Each curve represents a different amount of overhead. Fraud is the minority class. As expected, the larger overhead leads to higher cost. More importantly, we observe that when the overhead is smaller, the cost minimizes at a larger percentage of fraudulent transactions (minority class) in the training set. When the overhead is smaller, the bank can afford to send a larger number of transactions for investigation. That is, the bank can tolerate more false-alarms (a higher false-positive rate) and aim for fewer misses (a lower false-negative rate), which can be achieved by a larger percentage of fraudulent transactions (positive's). Conversely, if the overhead is larger, the bank should aim for fewer false-alarms (a lower FP rate) and tolerate more misses (a higher FN rate), which can be obtained by a smaller percentage of positive's. (Note that, at some point, the overhead can be large enough making fraud detection economically undesirable.)

The test set (from 9/96) has 40,038 transactions and 17.5% of them are fraudulent. If fraud detection is not available, on the average, $36.96 is lost per transaction. Table 1 shows the maximum savings of each algorithm with the most effective fraud percentage during training. *Cost* is the dollars lost per transaction; *fraud%* denotes the most effective fraud percentage for training; *%saved* represents the percentage of savings from the average loss of $36.96; *$saved* shows the total dollars saved for the month (9/96). BAYES performed relatively poor (we suspect the way we are treating attributes with real values in BAYES is not appropriate for the fraud domain) and is excluded from the following discussion. Considering the amount of overhead ranged

Table 1: Cost and saving in the credit card fraud domain

Learning Alg.	Overhead = $50				Overhead = $75				Overhead = $100			
	Cost	fraud%	%saved	$saved	Cost	fraud%	%saved	$saved	Cost	fraud%	%saved	$saved
C4.5	23.85	50%	35%	525K	26.88	30%	27%	404K	28.46	30%	23%	341K
CART	20.80	50%	44%	647K	23.64	50%	36%	534K	26.05	50%	30%	437K
RIPPER	21.16	50%	43%	632K	24.23	50%	34%	510K	26.73	50%	28%	409K
BAYES	35.23	30%	5%	69K	35.99	20%	3%	39K	36.58	20%	1%	15K

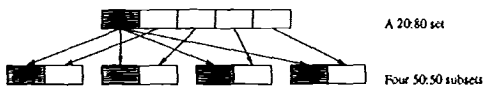

Figure 2: Generating four 50:50 data subsets from a 20:80 data set

from $50 to $100, the learning algorithms we used generally achieved at least 20% in savings or at least $300K. With an overhead of $75 or less, at least half a million dollars in savings can be attained. More importantly, maximum savings were achieved 7 out of 9 times when the fraud percentage used in training is 50%, which is not the "natural" 20% in the original data set.

In summary, based on our empirical results in this study, we observe that varying the training class distribution affects the cost performance. Since the "natural" distribution is 20:80, one way to achieve the "desired" 50:50 distribution is to ignore 75% of the legitimate transactions (or 60% of all the transactions), as we did in the experiments above. The following section discusses an approach that utilizes all the data and improves cost performance and scalability.

A Multi-classifier Meta-learning Approach to Non-uniform Distributions

As we discussed earlier, using the natural class distribution might not yield the most effective classifiers (particularly when the distribution is highly skewed). Given a skewed distribution, we would like to generate the desired distribution without removing any data. Our approach is to create data subsets with the desired distribution, generate classifiers from the subsets, and integrate them by learning (meta-learning) from their classification behavior. In our fraud domain, the natural skewed distribution is 20:80 and the desired distribution is 50:50. We randomly divide the majority instances into 4 partitions and 4 data subsets are formed by merging the minority instances with each of the 4 partitions containing majority instances. That is, the minority instances are replicated across 4 data subsets to generate the desired 50:50 distribution. Figure 2 depicts this process.

Formally, let n be the size of the data set with a distribution of $x : y$ (x is the percentage of the minority class) and $u : v$ be the desired distribution. The number of minority instances is $n \times x$ and the desired number of majority instances in a subset is $nx \times \frac{v}{u}$. The number of subsets is the number of majority instances ($n \times y$) divided by the number of desired majority instances in each subset, which is $\frac{ny}{\frac{nxv}{u}}$ or $\frac{y}{x} \times \frac{u}{v}$. (When it is not a whole number, we take the ceiling ($\lceil \frac{y}{x} \times \frac{u}{v} \rceil$) and replicate some majority instances to ensure all of the majority instances are in the subsets.) That is, we have $\frac{y}{x} \times \frac{u}{v}$ subsets, each of which has nx minority instances and $\frac{nxv}{u}$ majority instances.

The next step is to apply a learning algorithm(s) to each of the subsets. Since the subsets are independent, the learning process for each subset can be run in parallel on different processors. For massive amounts of data, substantial improvement in speed can be achieved for super-linear-time learning algorithms.

The generated classifiers are combined by learning (meta-learning) from their classification behavior. Several meta-learning strategies are described in (Chan & Stolfo 1993). To simplify our discussion, we only describe the *class-combiner* (or *stacking* (Wolpert 1992)) strategy. In this strategy a meta-level training set is composed by using the (base) classifiers' predictions on a validation set as attribute values and the actual classification as the class label. This training set is then used to train a meta-classifier. For integrating subsets, class-combiner can be more effective than the voting-based techniques (Chan & Stolfo 1995).

Experiments and Results

To evaluate our multi-classifier meta-learning approach to skewed class distributions, we used transactions from the first 8 months (10/95 – 5/96) for training, the ninth month (6/96) for validating, and the twelfth month (9/96) for testing (the two-month gap is chosen according to the amount of time needed to completely determine the legitimacy of transactions and simulates real life). Based on the empirical results in the previous section, the desired distribution is 50:50. Since the natural distribution is 20:80, four subsets are generated from each month for a total of 32 subsets. We applied four learning algorithms (C4.5, CART, RIPPER, and BAYES) to each subset and generated 128 base classifiers. BAYES (more effective for meta-learning in our experience) was used to train the meta-classifier.

Results from different amounts of overhead are plotted in Figure 3. Each data point is the average of ten runs using different random seeds. To demonstrate that 50:50 is indeed the desired distribution, we also performed experiments on other distributions and plotted the results in the figure. As expected, the cost is minimized when the fraud percentage is 50%. Surprisingly,

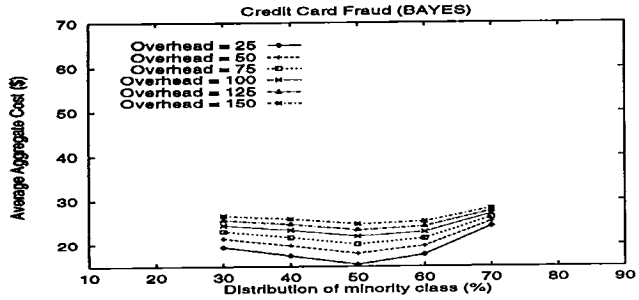

Figure 3: Training distribution vs. cost using meta-learning

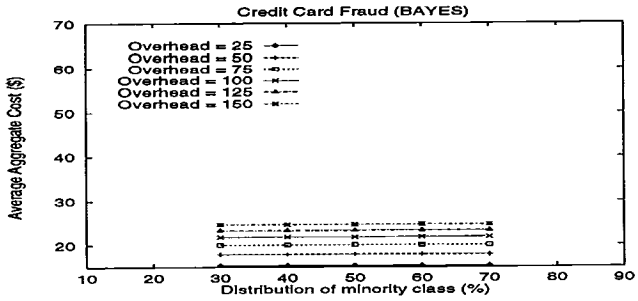

Figure 4: Distribution in validation set vs. cost using meta-learning

50% is the desired distribution for any of the overhead amount. This is different from the results obtained from previous experiments when meta-learning was not used.

Furthermore, to investigate if our approach is indeed fruitful, we ran experiments on the class-combiner strategy directly applied to the original data sets from the first 8 months (i.e., they have the natural 20:80 distribution). We also evaluate how individual classifiers generated from each month perform without class-combining. Table 2 shows the cost and savings from class-combiner using the 50:50 distribution (128 base classifiers), the average of individual CART classifiers generated using the desired distribution (from Table 1), class-combiner using the natural distribution (32 base classifiers—8 months × 4 learning algorithms), and the average of individual classifiers using the natural distribution (the average of 32 classifiers). (We did not perform experiments on simply replicating the minority instances to achieve 50:50 in one single data set because this approach increases the training set size and is not appropriate in domains with large amounts of data—one of the three primary issues we try to address here.) Compared to the other three methods, class-combining on subsets with a 50:50 fraud distribution clearly achieves a significant increase in savings—at least $110K for the month (6/96). When the overhead is $50, more than half of the losses were prevented. Surprisingly, we also observe that when the overhead is $50, a classifier ("single CART") trained from one's month data with the desired 50:50 distribution (generated by throwing away some data) achieved significantly more savings than combining classifiers trained from all eight months' data with the natural distribution. This reaffirms the importance of employing the appropriate training class distribution in this domain.

Class distribution in the validation set

Thus far we have concentrated on the class distributions in training the base classifiers. We hypothesize that the class distribution in the validation set (and hence the meta-level training set) affects the overall performance of meta-learning. To investigate that hypothesis, in this set of experiments, we fixed the training distribution of the base classifiers to 50:50 and varied the distribution of the validation set from 30:70 to 70:30 (as in the previous set of experiments). The different distributions are generated by keeping all the fraudulent transactions and randomly selecting the corresponding number of legitimate transactions (i.e., some legitimate records are thrown away).

Each data point in Figure 4 is an average of 10 runs using different random seeds (though the same seeds as in previous experiments). To prevent amplifying small differences unnecessarily, we use the same scale in Figure 4 as in the other similar figures. Unexpectedly, the curves are quite flat. That is, changing the class distribution in the validation set does not seem to affect the cost performance (a subject of further investigation).

Table 3 has the same format as the previous cost and savings tables, but the *fraud%* columns denote the fraud percentage in the validation set. The first row shows the results from class-combiner using the natural distribution in the validation set (from Table 2). The second row has the lowest cost performance and the corresponding fraud percentage in the validation set. The cost is minimized at around 50% (the difference is negligible among different distributions (Figure 4)), but it is not much different from the performance obtained from the natural distribution. Note that a 50:50 distribution was obtained by throwing away 60% of the data from the natural 20:80 distribution in the validation set. That is, comparable performance can be achieved with less data in the validation set, which is particularly beneficial in this domain that has large amounts of data.

Concluding Remarks

This study demonstrates that the training class distribution affects the performance of the learned classifiers and the natural distribution can be different from the desired training distribution that maximizes performance. Moreover, our empirical results indicate that our multi-classifier meta-learning approach using a 50:50 distribution in the data subsets for training can significantly reduce the amount of loss due to illegitimate transactions. The subsets are independent and can be processed in parallel. Training time can further be reduced by also using a 50:50 distribution in the validation set without degrading the cost performance. That is, this approach provides a means for efficiently

Table 2: Cost and savings using meta-learning

Method	Overhead = $50				Overhead = $75				Overhead = $100			
	Cost	fraud%	%saved	$saved	Cost	fraud%	%saved	$saved	Cost	fraud%	%saved	$saved
Class-combiner	17.96	50%	51%	761K	20.07	50%	46%	676K	21.87	50%	41%	604K
Single CART	20.80	50%	44%	647K	23.64	50%	36%	534K	26.05	50%	30%	437K
Class-combiner (nat.)	22.61	natural	39%	575K	23.99	natural	35%	519K	25.20	natural	32%	471K
Avg. single classifier	27.97	natural	24%	360K	29.08	natural	21%	315K	30.02	natural	19%	278K

Table 3: Cost and savings using different validation distributions (base distribution is 50:50)

Method	Overhead = $50				Overhead = $75				Overhead = $100			
	Cost	fraud%	%saved	$saved	Cost	fraud%	%saved	$saved	Cost	fraud%	%saved	$saved
Class-combiner (nat.)	17.96	natural	51%	761K	20.07	natural	46%	676K	21.87	natural	41%	604K
Class-combiner	17.84	70%	52%	766K	19.99	50%	46%	679K	21.81	50%	41%	607K

handling learning tasks with skewed class distributions, non-uniform cost per error, and large amounts of data. Not only is our method efficient, it is also scalable to larger amounts of data.

Although downsampling instances of the majority class is not new for handling skewed distributions (Breiman *et al.* 1984), our approach does not discard any data, allows parallelism for processing large amounts of data efficiently, and permits the usage of multiple "off-the-shelf" learning algorithms to increase diversity among the learned classifiers. Furthermore, how the data are sampled is based on the cost model, which might dictate downsampling instances of the minority class instead of the majority class.

One limitation of our approach is the need of running preliminary experiments to determine the desired distribution based on a defined cost model. This process can be automated but it is unavoidable since the desired distribution is highly dependent on the cost model and the learning algorithm.

Using four learning algorithms, our approach generates 128 classifiers from a 50:50 class distribution and eight months of data. We might not need to keep all 128 classifiers since some of them could be highly correlated and hence redundant. Also, more classifiers are generated when the data set is more skewed or additional learning algorithms are incorporated. Metrics for analyzing an ensemble of classifiers (e.g., diversity, correlated error, and coverage) can be used in pruning unnecessary classifiers. Furthermore, the real distribution is more skewed than the 20:80 provided to us. We intend to investigate our approach with more skewed distributions. As with a large overhead, a highly skewed distribution can render fraud detection economically undesirable. More importantly, since thieves also learn and fraud patterns evolve over time, some classifiers are more relevant than others at a particular time. Therefore, an adaptive classifier selection method is essential. Unlike a monolithic approach of learning one classifier using incremental learning, our modular multi-classifier approach facilitates adaptation over time and removal of out-of-date knowledge.

Acknowledgements This work was partially funded by grants from DARPA (F30602-96-1-0311), NSF (IRI-96-32225 & CDA-96-25374), and NYSSTF (423115-445). We thank Dave Fan and Andreas Prodromidis for their discussion and the reviewers for their comments.

References

Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 1984. *Classification and Regression Trees.* Belmont, CA: Wadsworth.

Cardie, C., and Howe, N. 1997. Improving minority class prediction using case-specific feature weights. In *Proc. 14th Intl. Conf. Mach. Learning*, 57–65.

Catlett, J. 1991. Megainduction: A test flight. In *Proc. Eighth Intl. Work. Machine Learning*, 596–599.

Chan, P., and Stolfo, S. 1993. Meta-learning for multistrategy and parallel learning. In *Proc. Second Intl. Work. Multistrategy Learning*, 150–165.

Chan, P., and Stolfo, S. 1995. A comparative evaluation of voting and meta-learning on partitioned data. In *Proc. Twelfth Intl. Conf. Machine Learning*, 90–98.

Clark, P., and Niblett, T. 1989. The CN2 induction algorithm. *Machine Learning* 3:261–285.

Cohen, W. 1995. Fast effective rule induction. In *Proc. 12th Intl. Conf. Machine Learning*, 115–123.

Fawcett, T., and Provost, F. 1997. Adaptive fraud detection. *Data Mining and Knowledge Discovery* 1:291–316.

Fawcett, T. 1996. Learning with skewed class distributions–summary of responses. Machine Learning List, Vol. 8, No. 20.

Kubat, M., and Matwin, S. 1997. Addressing the curse of imbalanaced training sets: One sided selection. In *Proc. 14th Intl. Conf. Machine Learning*, 179–186.

Pazzani, M.; Merz, C.; Murphy, P.; Ali, K.; Hume, T.; and Brunk, C. 1994. Reducing misclassification costs. In *Proc. 11th Intl. Conf. Machine Learning*, 217–225.

Provost, F., and Aronis, J. 1996. Scaling up inductive learning with massive parallelism. *Machine Learning* 23:33–46.

Quinlan, J. R. 1993. *C4.5: programs for machine learning.* San Mateo, CA: Morgan Kaufmann.

Tan, M. 1993. Cost-sensitive learning of classification knowledge and its applications in robotics. *Machine Learning* 13:7.

Wolpert, D. 1992. Stacked generalization. *Neural Networks* 5:241–259.