Sample Size and Modeling Accuracy with Decision Tree Based Data Mining Tools

By

James Morgan, Robert Dougherty, Allan Hilchie, and Bern Carey

All of the Center for Data Insight, Northern Arizona University

SAMPLE SIZE AND MODELING ACCURACY

WITH DECISION-TREE BASED DATA MINING TOOLS

ABSTRACT

Given the cost associated with modeling very large datasets and over-fitting issues of decision-tree based models, sample based models are an attractive alternative – provided that the sample based models have a predictive accuracy approximating that of models based on all available data. This paper presents results of sets of decision-tree models generated across progressive sets of sample sizes. The models were applied to two sets of actual client data using each of six prominent commercial data mining tools.

The results suggest that model accuracy improves at a decreasing rate with increasing sample size. When a power curve was fitted to accuracy estimates across various sample sizes, more than 80 percent of the time accuracy within 0.5 percent of the expected terminal (accuracy of a theoretical infinite sample) was achieved by the time the sample size reached 10,000 records. Based on these results, fitting a power curve to progressive samples and using it to establish an appropriate sample size appears to be a promising mechanism to support sample based modeling for a large dataset.

INTRODUCTION

Data mining has emerged as a practical analytical tool primarily on the basis of its ability to deal with the large volume of data available from databases and data warehouses. Rapid increases in processor speed coupled with continuing decreases in the cost of mass storage devices and other computer hardware have made it practical to collect and maintain massive databases. Data mining software is viewed as a tool that can perform undirected or semi-directed analysis and, thus, can be applied to the full length and width of very large data sets at much lower costs than analytical techniques requiring stronger human direction. While there is an inherent bias toward applying data mining tools to the full volume of available data, issues of cost and model over-fitting suggest that use of data mining models based on a sample of

available data may be appropriate in many instances. Thus, the relationship between sample size and model accuracy is an important issue for data mining.

Despite the increases in processing speeds and reductions in processing cost, applying data mining tools to analyze all available data is costly in terms of both dollars and time required to generate and implement models. In discussing differences between statistical and data mining approaches, Mannila [2000] suggests that: "The volume of the data is probably not a very important difference: the number of variables or attributes often has a much more profound impact on the applicable analysis methods. For example, data mining has tackled width problems such as what to do in situations where the number of variables is so large that looking at all pairs of variables is computationally infeasible." The above quote suggests that the benefit of data mining tools comes from their ability to deal more effectively with complex interactions among variables rather than from the ability to process massive volumes of instances.

It has been noted that decision tree based data mining tools are subject to over-fitting as the size of the data set increases, Domingos [1998] and Oates and Jensen [1997]. As Oates and Jensen [1998] note, "Increasing the amount of data used to build a model often results in a linear increase in model size, even when that additional complexity results in no significant increase in model accuracy." In a similar vein Musick, Catlett, and Russell [1993] suggest that "often the economically rational decision is to use only a subset of the available data." A variety of pruning algorithms have been proposed to deal with this problem, and most commercial data mining software using decision tree based algorithms incorporate the use of pruning algorithms. While pruning helps to limit the proportion by which model complexity increases as the amount of data increases, its effectiveness can only be assessed by examining the responsiveness of model complexity and model accuracy to changes in data set size.

Sampling can also be used as a tool to lower the cost of maintaining data mining based operational models. Lee, Cheung, and Kao [1998] have proposed a dynamic sampling technique to test for changes in a dataset. Their technique suggests using a sample of data to detect when enough change has occurred in the structure of a dataset to justify re-estimation of a model using the full set of available data. In addition to this monitoring role, periodic re-estimation of a decision tree model using a moderate sized sample of data may be the most cost effective way to maintain a reliable predictive model. For example, an organization might find it equally costly

to re-analyze a model on the basis of a sample of 10,000 records once a month or to re-analyze the model based on all available data once a year. In such a case, modeling based on a sample will be the most effective strategy if the phenomenon being modeled is relatively dynamic and models based on the sample approximate the accuracy of a model based on all available data.

Prior studies of sampling in data mining have used public domain data modeling tools and relatively small data sets from the UCI repository. In this paper we describe the results of models generated from the systematic sampling of data from two corporate datasets one of which contained more than 1.5 million records. The target variable for each data set is a binary variable. Models are generated with each of six prominent commercial data mining tools. Statistical analyses across the tools, over varying sample sizes, and with respect to other relevant factors are presented. The results provide an insight into the response of model accuracy with respect to increases in sample size, and also allow us to examine the extent to which that response varies across different data mining tools and across varied data sets.

REVIEW OF PREVIOUS SAMPLING STUDIES

The effectiveness of data mining models based on sampling from datasets has not been widely studied. However, there are a few studies that have addressed this topic which can be used as the starting point for this study.

John and Langley [1996] applied arithmetic progressive sampling (e.g. samples of 100, 200, 300, 400, etc.) to 11 of the UCI repository datasets. Because many of the datasets used were small, they first replicated each record 100 times to simulate a large dataset. The inflated data set was used to generate a set of samples whose size was systematically incremented by 100 records between samples. A model was then generated for each sample using a "naive Bayesian classifier." The sample-based models were applied to a holdout set to evaluate their accuracy. A power function based regression equation was estimated as each progressive sample was performed, and sampling was terminated when the accuracy of the current model was within 2 percent of the expected accuracy (based on the regression) for a model using the full dataset. Twenty-five sets of samples and their associated models were produced and tested for each dataset.

Applying this criterion to the 11 inflated UCI repository databases, led to average final sample sizes ranging from 300 to 2,180 all of which were within 2 percent of the accuracy of a

naïve Bayesian classifier model built from the entire training set. Limitations of this study include the fact that the results were generated by replicating data from small source datasets and that the models that were compared used naïve Bayesian classifiers.

Frey and Fisher [1999] systematically examined the response of modeling accuracy to changes in sample size using the C4.5 decision tree algorithm applied to 14 datasets from the UCI repository. The datasets used were all relatively small – from 57 to 3,196 observations. This study focused on determining the shape of the learning curve and made no attempt to determine an optimal sample size. For 13 of the 14 datasets, they found that the response of predictive accuracy to sample size was more accurately predicted by a regression based on a power law function than by regressions using linear, logarithmic, or exponential functions. The power coefficient varied rather substantially across the datasets (from +.118 to –1.117) .

Provost, Jensen, and Oates [1999] modeled 3 of the larger (32,000 record CENSUS, 100,000 record LED, and 100,000 record Waveform) UCI repository datasets using differing progressive sampling techniques. Progressive sampling begins with a relatively small sample from the dataset. Next, a model is created and run against a holdout dataset to test its accuracy. Then a larger sample is used to generate another model whose accuracy also is tested on the holdout set. The process is repeated for models based on progressively larger samples, until some standard accuracy criteria is met.

The primary aim of their paper was to compare the efficiency of alternative progressive sampling techniques as measured by the computation time required to achieve a standard degree of accuracy. Arithmetic, geometric, and dynamic progressive sampling techniques were evaluated. Arithmetic progressive sampling uses equal absolute increments between samples. For example, increments of 100 (100, 200, 300, 400) or increments of 500 (500, 1,000, 1,500, 2,000). Geometric progressive sampling uses equal proportional increments and an arbitrary initial size. For example, incremental doubling with an initial sample size of 100 would use samples of 100, 200, 400, 800. The dynamic progressive sampling technique used by Provost, Jensen, and Oates involved: (1) initially estimating and testing models based on samples of 100, 200, 300, 400, and 500, (2) estimating a power function based learning curve based on results for those models, and (3) selecting the next sample to be the size required to achieve the accuracy criteria according to the learning curve.

The initial accuracy criteria used called for sampling to progress until the average accuracy for the set of the last 3 samples is no more than 1% less accurate than results from a model based on all available data. At that point, the middle sample of the set is designated as the minimum sample meeting the criterion. This criterion was applied to an arithmetic sampling schedule with increments of 1,000. The criterion was met at the level of 2,000 records for the LED dataset, at 8,000 for the CENSUS dataset, and at 12,000 for the WAVEFORM dataset. Since this measure compares sample based models to the accuracy of a model based on the full dataset, it is clearly designed as a test of how accurate models based on various sample sizes are rather than as a method for determining what sample size is sufficient for a dataset whose population has not been modeled.

PLAN OF THE CURRENT STUDY

Our study incorporates sampling structures and evaluation techniques used in prior studies, but applies these techniques to real client data sets and models constructed using alternative commercial data mining tools.

For each data set to be analyzed, a holdout set of records was first removed and then a geometric progression of sample sizes was generated from the remaining training data set. The samples start at a size of 500 records and double for each new sample up to final a sample size of 32000, resulting in sample sizes of 500, 1,000, 2,000, 4,000, 8,000, 16,000, and 32,000. AN RS/6000 Sp/2 system provided to the CDI by IBM was used fro data preparation. For each sample size, a set of four distinct samples was generated with replacement. A model was created for each sample at each size using each of six data mining software tools. The tool used included: *name1, name2, name3, name4, name5, name6.*

The staff of the Center for Data Insight includes student workers responsible for mastering the use of a number of commercial data mining tools supplied by vendor partners. The analyses presented here compare results obtained using decision tree models from the six different commercial data mining tools, and built by the student expert on each tool. Nondisclosure agreements prevent us from identifying the tools in the comparative analyses – they are labeled as tool A through tool F in the analyses presented here.

Our goal has been to apply the sampling structure described above to a variety of data sets associated with “real” business customers of the Center. Initially we present results for two

data sets with binary target variables relating to differing aspects of customer relationship management issues. These data sets and target variables are briefly described below.

Dataset 1 consists of data from a company selling computer related products largely to wholesale customers. The target variable for this dataset is a binary flag indicating whether a customer is still "active" or is a "dead" customer based on an appropriate criteria related to the recency of their last purchase. The proportion of active customers was approximately two-thirds. This dataset is relatively narrow (15 explanatory variables) with several of the explanatory variables containing cardinal numeric values. The full dataset includes approximately 50,000 customer records. Because of the relatively small size of dataset 1, a holdout set of 10,000 records was used for testing the models.

Dataset 2 tracks retail customers of a firm selling a broad range of products. The target variable is a binary variable classifying customers as "better than average" or "poorer than average." Customers were considered better than average if their score based on a weighted average of a set of measures of customer value was higher than average.[1] Thus, the target variable is evenly balanced between "better than average" and "poorer than average" customers. This dataset is over 100 variables in width, most of the explanatory variables are categorical, and the full dataset includes about 1.5 million customer records. A holdout set of 50,000 records was used for testing the models of the second dataset.

Since the datasets described above were from prior customers of the Center, the student modelers were reasonably familiar with the data. The modelers were encouraged to treat the study as a contest and build the most accurate model possible for their tool using any pruning parameters or other modeling options they felt appropriate. However, they were told to use a common model across all samples. Thus, they applied modeling options based on their apriori experience with the tool and maintained consistent modeling options across the various samples for a given dataset. The proportion of records correctly classified was used as the criteria for measuring the success of models.

[1] The measure used was a weighting of the recency of last purchase, frequency of purchases, and monetary value of purchases.

SUMMARY MODEL RESULTS

In examining model results, we will first look at summary measures comparing the performance of each tool at various sample sizes for each dataset. Table 1 and Figures 1 and 2 present averages (across the four samples) of the percentage of cases correctly classified for each tool at each sample size. In general, accuracy tends to increase at a decreasing rate as sample size increases. For dataset 1, tool B performed substantially less well than the others for all sample sizes below 16,000. The remaining 5 tools show relatively stable patterns with accuracy increasing at a decreasing rate. For all of the tools, model accuracy increases only modestly beyond the 16,000 sample size. For dataset 2, all of the tools produced rather smooth curves with accuracy increasing at a decreasing rate and becoming relatively flat for sample sizes of 8,000 or more.

Table 1
Average Percentage Correctly Classified by Tool and Dataset

Sample Size	Tool A	Tool B	Tool C	Tool D	Tool E	Tool F
Dataset 1						
500	82.07	71.82	80.97	80.28	83.41	83.84
1,000	83.88	67.87	83.06	81.18	85.17	83.62
2,000	84.23	66.29	83.48	82.70	85.88	84.60
4,000	84.92	68.96	85.85	83.15	86.31	85.94
8,000	85.24	69.91	85.36	82.98	86.45	82.78
16,000	85.39	85.80	85.84	86.88	86.21	86.48
32,000	85.23	85.80	86.11	87.70	86.51	86.78
Dataset 2						
500	86.04	86.78	90.35	82.40	87.41	89.70
1,000	87.38	89.16	93.58	88.33	88.60	89.91
2,000	88.08	90.36	93.19	88.65	89.63	91.19
4,000	90.33	91.45	93.45	89.85	90.63	91.96
8,000	89.92	91.94	93.63	90.10	91.42	92.40
16,000	90.34	92.57	93.07	90.23	91.56	92.87
32,000	90.60	93.11	93.95	90.55	91.61	93.34

Figure 1

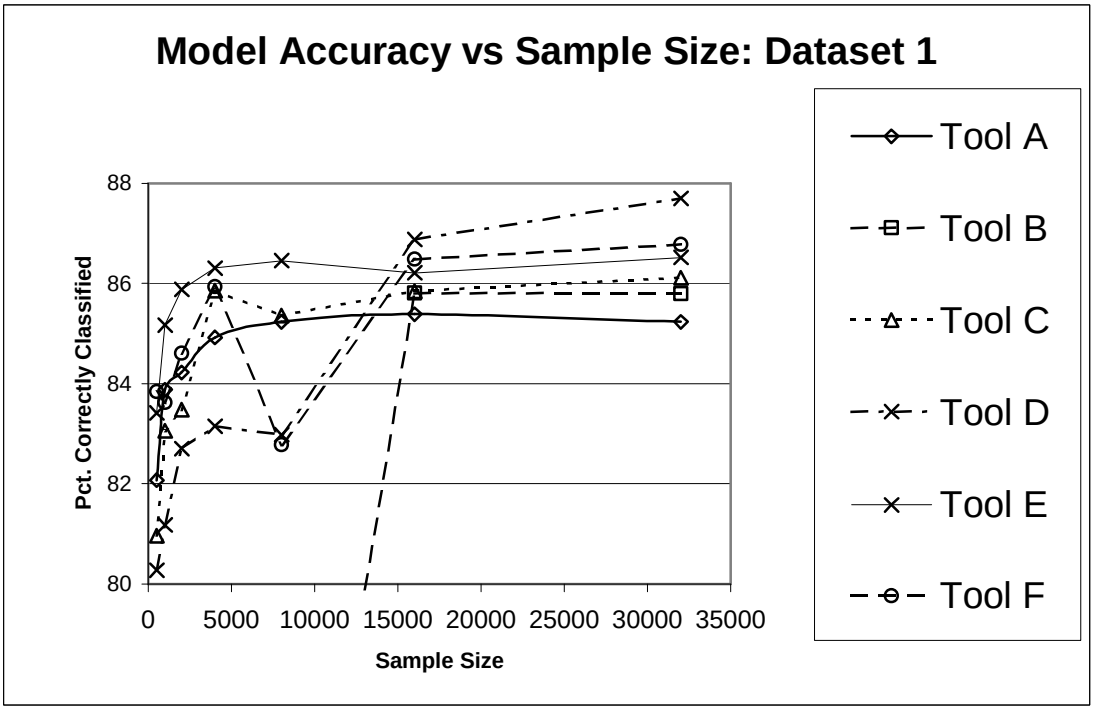
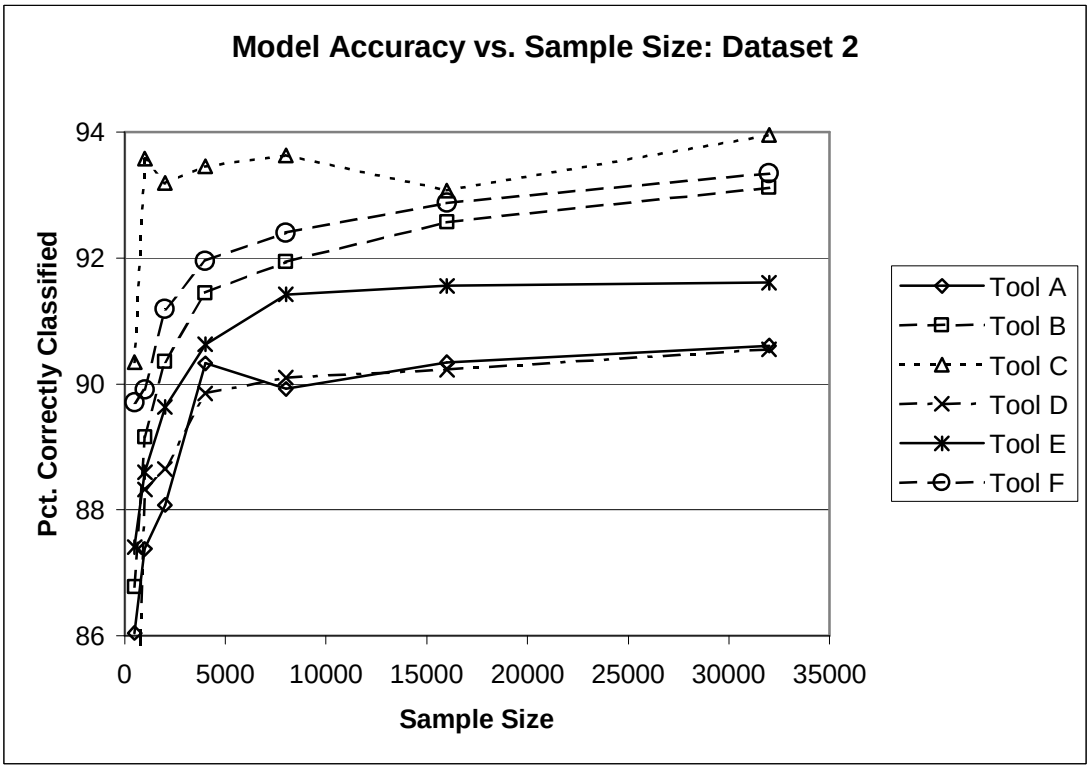


Figure 2



Within a given dataset tools that do well for small sample sizes also tended to do well for larger sizes, however, there was no clear pattern of dominance across the two datasets. Tool E provided the most consistent performance on dataset 1, while tool D showed the best performance at the largest sample sizes. Both of these tools were near the bottom in their performance for dataset 2. At the same time, tool C provided the strongest accuracy for dataset 2 (particularly at small sample sizes), but had average performance for dataset 1.

Table 2 presents summary univariate analysis of variance results for the two datasets. The sample size and the tool used as well as the interaction between those 2 factors were used as explanatory factors. As Table 2 indicates, both factors and their interaction are statistically significant for both datasets.

Also presented are results of the Tukey test for homogeneous subsets for each factor. This test identifies sets of values for each factor whose means do not differ significantly at the .05 level. Results for both datasets show that accuracy increases with sample size. Results for dataset 2 show a plateauing of accuracy – accuracy for all sample sizes above 4,000 fit into the same homogeneous subset. Dataset 1 results place only the 16,000 and 32,000 sample sizes in the final homogeneous subset. This is primarily due to the unusual pattern of results for Tool B. If tool B is excluded for dataset 1, all sample sizes greater than 4,000 once again fit in the same homogeneous subset. The Tukey test for the tool factor shows that there are significant differences in average accuracy, but no strong systematic patterns that hold up across both datasets.

Finding an Optimal Sample Size

In the previous section all of the data presented was based on average responses across the four separate samples that were generated for each sample size. While this information is of interest, the robustness of the results across individual samples is perhaps of more interest.

For someone contemplating using sample data for modeling, knowing that 90 percent of the time a sample of 8,000 records will be accurate to within 0.5 percent of the accuracy of a model based on all available data is likely to be more useful than knowing that the average sample of 8,000 records is only .25 percent less accurate than a model based on all available data. Sampling will be accepted if there is only a small probability that its accuracy will be outside of acceptable bounds. In addition, it is likely that the sample size required to approach the accuracy of a model

based on all available data will vary considerably from one dataset to another. This suggests that sampling should ideally be approached on a dataset-by-dataset basis. Under this scheme, a number of progressive samples at relatively small sizes would be used to build models. Some measurement designed to test for convergence in the accuracy of the models would then be applied to determine whether additional larger samples were needed to achieve acceptable model accuracy.

Table 2

Univariate Analysis of Variance Results

	Dataset1		Dataset 2	
Adjusted R-Squared	*0.891*		*0.753*	
Source	**F Value**	**Signif.**	**F Value**	**Signif.**
Corrected Model	34.31	*0.000*	13.44	*0.000*
Intercept	332167.86	*0.000*	818558.80	*0.000*
Size	42.05	*0.000*	44.73	*0.000*
Tool	161.22	*0.000*	46.37	*0.000*
Size*Tool	11.61	*0.000*	1.69	*0.024*

Tukey Test for Homogeneous Subsets

Dataset 1

	Homogeneous Subset				Homogeneous Subset			
Sample Size	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
500	80.40				87.12			
1000	80.80	80.80				89.49		
2000	81.19	81.19	81.19			90.18	90.18	
4000			82.52				91.28	91.28
8000		82.12	82.12					91.57
16000				86.10				91.77
32000				86.36				92.19
Signif.	*0.755*	*0.174*	*0.170*	*0.999*		*0.517*	*0.054*	*0.178*

	Homogeneous Subset			Homogeneous Subset			
Tool Used	**1**	**2**	**3**	**1**	**2**	**3**	**4**
Tool A		84.42	84.42	88.96			
Tool B	73.78				90.77	90.77	
Tool C		84.38	84.38				93.03
Tool D		83.55		88.59			
Tool E			85.71		90.12		
Tool F		84.86	84.86			91.62	
Signif.		*0.089*	*0.083*	*0.894*	*0.428*	*0.133*	

The models created in this study followed a fixed sampling scheme using progressive samples form 500 to 32,000 records. However, tests of progressive sampling methodologies can be applied ex-post. We can evaluate the sample size that would have been required to meet a particular convergence criterion. Two alternative methods for measuring convergence across progressive samples are presented here. The first is based on using moving averages of model accuracy while the second uses statistical analysis of model results.

Oates and Jensen [1997] used a convergence measure based on examining the improvement in model accuracy as the sample size was progressively increased to test for convergence. Under this scheme, when the improvement in accuracy drops below a prescribed level sampling is terminated. For this criterion to be effective, the improvement in model accuracy as sample size increases needs to be relatively stable and monotonically decreasing in magnitude. If this is the case, it is reasonable to assume that, once the improvement in model accuracy drops below a specified limit, it would stay below that limit for all larger sample sizes as well, and that a plateau in model accuracy has been achieved. To minimize the chance of improperly terminating due to a single non-representative sample, a moving average of the accuracy of the last three samples is maintained and sampling is terminated when the improvement in this moving average drops below a specified convergence criterion (1 percent in Oates and Jensen's paper).

In adapting this technique, we used a weighted average of the last 3 samples. That is, the moving average model accuracy for a given sample sizes is:

$$AccMA_n = (Sz_n*Acc_n + Sz_{n-1}*Acc_{n-1} + Sz_{n-2}*Acc_{n-2})/(Sz_n + Sz_{n-1} + Sz_{n-2})$$

where Acc_n is the measured model accuracy for the nth progressive sample, Sz_n is the size of the nth progressive sample, and $AccMA_n$ is the moving average accuracy measure for the nth progressive sample. The convergence test applied calls for sampling to terminate if

$$AccMA_n - AccMA_{n-1} < ?$$

where ? is the convergence criterion. For this study, values of both 1 percent and 0.5 percent are used for ?.

Summary results for this convergence criterion (applied to the models generated from each of the four sample-sets for each tool across the two datasets) are presented in Table 3. When a 1 percent convergence criterion is used, convergence is achieved by the time a sample

size of 8,000 records is reached in almost every instance across both datasets. When the 0.5 percent criterion is used, there is more variety in the sample size required. However, three-quarters of the sample-set/tool combinations for dataset 1, and over 60 percent of the sample-set/ tool combinations for dataset 2 reached convergence at 8,000 records or less.

Table 3

Moving Average Convergence Results for Sampling Runs

	Using 1% Convergence Criterion				Using 0.5% Convergence Critireon			
	Dataset1		Dataset 2		Dataset 1		Dataset 2	
Sample size at Convergence	Number	Pct.	Number	Pct.	Number	Pct.	Number	Pct.
4,000	14	58.3%	11	45.8%	7	29.2%	5	20.8%
8,000	9	37.5%	13	54.2%	11	45.8%	10	41.7%
16,000	0	0.0%	0	0.0%	3	12.5%	3	12.5%
32,000	0	0.0%	0	0.0%	0	0.0%	4	16.7%
> 32,000	1	4.2%	0	0.0%	3	12.5%	2	8.3%
Unstable*	8	33.3%	2	8.3%	10	41.7%	6	25.0%

* A set of samples is considered unstable if the convergence criterion is met at one sample size but is not met for some larger sample size.

Since this analysis was performed ex-post, we were able to test the stability of sample-sets meeting the convergence criteria at sample sizes less than 32,000. Moving average accuracy values for each sample size up to 32,000 were always computed. If a sample-set meeting the convergence criterion for one sample size would have failed that test at some larger sample size it was classified as unstable. For example, if sample-set 2 for tool C met the 1 percent convergence criterion at a sample size of 8,000, we would look at the change in moving average accuracy from 8,000 to 16,000 and from 16,000 to 32,000. If either of these showed an improvement of more than 1 percent, the model would be classified as unstable for that sample-set. The results of Table 3 show only 2 of 24 models to be unstable for dataset 2 with the convergence criterion set at 1 percent. However one-third of the sample-set/tool combinations show unstable results for dataset 1. When the convergence criterion is tightened to 0.5 percent, unstable results are found for one-quarter of the sample-set tool combinations of dataset 2 and over 40 percent of those for dataset 1.

While the moving average results are interesting, the number of exceptions found is somewhat troubling. Also, there is no means of estimating just how close a sample-based model's accuracy is to the accuracy that could be expected from a model using all available data. To provide such a convergence criterion we need to produce a model of the shape of the response of accuracy to changes in sample size that either provides an upper limit on accuracy as simple size increases or estimates a curve that can be extrapolated to estimate expected accuracy for the total number of records available in the dataset.

Casual observation of Figures 1 and 2 suggests a shape that is consistent with a log-linear model or a power curve model. Power curve models approach a finite limit while log-linear models are theoretically unbounded. Because the dependent variable in this study is the percentage of cases correctly classified, boundedness is an attractive property. In addition, Frey and Fisher's [1999] results cited earlier indicate that the power curve tends to provide a strong fit (stronger than linear, log-linear, or exponential models in 13 of the 14 datasets they modeled). For these reasons, a model based on the power curve was used in analyzing the response of accuracy to sample size.

The form of model used was:

$$\text{acc}(n) = a - be^{nc}$$

where n is the sample size, acc(n) is the expected accuracy of a model whose sample sizes is n, *a*, *b*, and *c* are parameters to be estimated, and e is the natural logarithm. For well-behaved systems the value of *b* is positive and the value of *c* is negative. When this is the case, the term be^{nc} approaches 0 as n becomes large. Thus, the value of *a* can be interpreted as an asymptotic value representing the accuracy that would be produced by the model with an infinitely sized dataset (hereafter terminal accuracy). The values of the *b* and *c* parameters interact in determining the shape of the response curve in a way that makes their direct interpretation somewhat difficult. It is of more interest to apply the model and obtain estimates of the sample size required to bring the expected model accuracy within a fixed percentage of the asymptotic accuracy.

Table 4 presents summary results of nonlinear regressions using this model for each sample-set across tools and datasets. Each model is based on all 7 sample sizes from 500 to 32,000. Given the complexity of the non-linear model to be estimated, generation of stable models for samples up to some smaller sample size is problematic. Even with all sample sizes

included, the models have only 3 degrees of freedom. R-Squared values are not shown, but generally suggest that the models are rather strong. Sixteen of the 24 models generated from dataset 1 had an R-squared greater than 0.9, while 13 of the models for dataset 2 met this criterion. The column labeled terminal accuracy presents the *a* parameter, the *estimated* terminal accuracy. In addition, the estimated sample sizes required to come within 1 percent and within 0.5 percent of this level of accuracy are also presented. In two instances, the *a* parameter was greater than 100 percent leading to an unstable model. Those instances are shown as the starred entries.

It is interesting to note the degree of consistency in the *a* parameter across sample-sets for each tool. For dataset 1, 3 of the tools had less than 1 percent variation in the *a* parameter across the four sample-sets, while 4 of the 6 tools met this criterion for dataset 2.

Table 4 also suggests that relatively small samples will often produce models whose accuracy approaches that of an unlimited sample size. For 22 of the 24 models from dataset 2, accuracy came within 0.5 percent of the terminal accuracy at a sample size less than 10,000. For dataset 1, models for tools B and D consistently approach their terminal accuracy only at a substantially higher sample size. Thus, only 15 of the 24 models based on dataset 1 came within 0.5 percent of their terminal accuracy at a sample size less than 10,000.

Overall, the results in Table 4 suggest that relatively small samples can often be used to build models whose accuracy approximates that of models built from the full set of available data. Also, these results are reasonably comparable to those of the Provost, Jensen, and Oates paper that found convergence to 1 percent at sample sizes between 2,000 and 12,000 for selected datasets in the UCI repository. However, the number of exceptions is somewhat troubling. Also, the systematic nature of the exceptions reinforces the idea that the sample size needed to approach terminal accuracy is likely to vary from dataset to dataset.

One could think of the models presented in Table 4 as a procedure to be applied in determining the sample size needed to adequately model a particular dataset. A progressive set of samples for all sizes up a certain limit would be modeled and a power curve estimated. If the power curve suggests that a larger sample is needed to come within a desired limit of terminal accuracy, an additional larger sample would be taken. Additional sampling might be continued on the same basis (double the last sample size used) and the power curve re-estimated until the

convergence criterion is met. Alternatively, one additional sample might be generated at the size required to reach the convergence criterion based on the initial power curve estimate.

Table 4

Sample Size Required to Achieve Convergence Across Alternative Tools and Datasets

		Dataset 1			Dataset 2		
			Sample to approach limit			Sample to approach limit	
Tool	Sample -Set	Terminal Accuracy	within 1 %	within 0.5 %	Terminal Accuracy	within 1 %	within 0.5 %
Tool A	1	85.14	1,942	3,083	90.54	4,921	6,778
	2	85.30	1,909	3,271	90.12	2,511	3,515
	3	85.25	2,018	2,761	91.12	4,721	8,199
	4	85.19	788	943	90.09	2,261	2,926
Tool B	1	96.67	95,216	114,669	92.63	3,548	4,739
	2	****	****	****	92.56	4,122	5,854
	3	91.05	57,361	69,785	92.33	3,387	4,810
	4	91.59	54,796	66,055	92.35	2,068	2,608
Tool C	1	85.85	3,565	5,020	94.43	14,541	25,877
	2	85.92	4,046	6,785	****	****	****
	3	85.51	1,249	1,503	93.55	750	875
	4	85.93	2,959	3,977	94.23	593	625
Tool D	1	87.11	10,146	13,244	89.92	1,551	1,859
	2	94.93	131,070	165,907	90.59	5,016	9,264
	3	89.30	35,466	47,547	89.92	986	1,119
	4	88.05	21,147	28,494	90.22	1,722	2,242
Tool E	1	86.61	1,549	2,155	91.63	3,656	5,742
	2	86.35	1,216	2,235	90.90	1,142	1,430
	3	86.36	1,179	1,700	91.50	3,046	4,331
	4	86.15	874	1,049	91.71	4,004	5,481
Tool F	1	****	****	****	93.08	5,743	9,088
	2	86.78	4,884	9,368	93.01	5,876	9,519
	3	86.69	4,871	8,137	92.90	3,323	4,804
	4	86.39	3,132	4,351	93.24	3,935	5,587

In our data, assuming that samples up to 32,000 were initially created and modeled, 40 of the 48 sample-set/tool combinations would meet the criterion of coming within 0.5 percent of terminal accuracy at or before the 32,000 sample size. Two of the remaining 8 sample-set/tool combinations did not produce a stable power curve, suggesting either that the full dataset be used for modeling or that the next progressive sample size should be applied and the power curve re-

estimated until a stable model meeting the convergence criterion is achieved. For the 6 sample-set/tool combinations whose convergence sample size was larger than 32,000, a new sample at the size required to meet the convergence criterion would be drawn and modeled using the appropriate tool (or the full dataset would be used if the dataset size is less than the sample size to meet the convergence criterion).

The usefulness of the approach outlined in the previous paragraph is evident for the data of dataset 2. For 23 of the 24 sample-set/tool combinations, a model whose expected accuracy is within 0.5 percent of the terminal accuracy was found by running a data mining tool against a total of 65,500 records. The computation time required for this would be substantially less than that required to model against the full 1.5 million record dataset.

Summary

This paper has presented the results of decision-tree models generated using systematic sets of progressive sample sizes. The analyses presented here were applied to 2 sets of actual client data using each of 6 prominent commercial data mining tools.

Comparisons of results across tools indicated significant differences in the effectiveness of the various tools in modeling particular datasets. However, there was not a consistent pattern of tool performance across the 2 datasets. The tools that performed best on dataset 1 were not particularly strong for dataset 2 and vice-versa.

In general, our results suggest that model accuracy tends to increase at a decreasing rate with increases in sample size. In most cases, the results were fit rather well by a model that assumes that the response of accuracy to increases in sample size can be specified by a power curve with a finite terminal value less than 100 percent. The power curve is characterized by a long plateau, with values close to the terminal value at large sample sizes. While rather erratic performance was observed for some of the small samples from dataset 1, accuracy almost universally reached a plateau by the time the 16,000 record sample size was reached. More than 80 percent of the time, accuracy within 0.5 percent of the expected terminal accuracy was achieved by the time the sample size reached 10,000 records. Results for dataset 2 were substantially more consistent than those for dataset 1, reinforcing the idea that the size of sample

needed to achieve adequate model performs is likely to vary substantially across dataset and target variable characteristics.

Our results do suggest that systematic progressive sampling often produces models whose expected accuracy is very close to the accuracy expected from a model based on the full dataset. Fitting a power curve to a set of progressive samples and using its results to assess the adequacy of the samples used and determine the appropriate size for an additional sample, if needed, appears to be a promising mechanism for sample-based mining of a large dataset.

This preliminary work suggests a number of avenues for further research. Examination of sampling responsiveness should be extended to broader types of datasets and to non-binary target variables and target variables whose distribution is skewed to varying degrees. Another interesting extension to this study would be the systematic application of bagging of the samples required to produce the accuracy responsiveness estimates, which might provide a low cost means to fully utilize all the samples required to apply this technique.

References

Domingos, P., 1998, "Occam's Two Razors: the Sharp and the Blunt," **Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining**, Menlo Park, CA: AAAI Press, pp. 37-43.

Frey, L. and Fisher D., 1999, "Modeling Decision Tree Performance with the Power Law," **Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics**, San Francisco, CA: Morgan-Kaufmann, pp59-65.

John, G. and Langley, P., 1996, "Static Versus Dynamic Sampling for Data Mining," **Proceedings of the Second International Conference on Knowledge Discovery and Data Mining**, , AAAI Press, pp. 367-370.

Lee, S., Cheung, D., and Kao, B., 1998, "Is Sampling Useful in Data Mining? A Case in the Maintenance of Discovered Association Rules," **Data Mining and Knowledge Discovery**, Vol. 2, Kluwer Academic Publishers, pp. 232-262.

Mannila, H., 2000, "Theoretical Frameworks for Data Mining," **SIGKDD Explorations**, Vol. 1, No. 2, ACM SIGKDD, pp. 30-32.

Musick, R., Catlett, J, and Russel, S., 1993, "Decision Theoretic Subsampling for Induction on Large Databases," **Proceedings of the Tenth International Conference on Machine Learning**, San Mateo, CA: Morgan Kaufmann, pp. 212-219.

Oates, T. and Jensen, D., 1997, "The Effects of Training Set Size on Decision Tree Complexity," **Machine Learning: Proceedings of the Fourteenth International Conference**, Morgan Kaufmann, pp. 254-262.

Oates, T. and Jensen, D., 1998, "Large Data Sets Lead to Overly Complex Models: an Explanation and a Solution," **Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining**, Menlo Park, CA: AAAI Press, pp. 294-298.

Provost, F., Jensen, D, and Oates, T., "Efficient Progressive Sampling", **Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining**, San Diego, CA: ACM SIGKDD, pp. 23-32.