

Static vs Dynamic Sampling for Data Mining

George H. John and Pat Langley
Computer Science Dept.,
Stanford University

Presented by: Ram Kolli

Problem

- There is a proliferation of huge data warehouses with terabytes of data.
- All kinds of decision support systems are dependent on mining information from this huge amount of data.
- There are a few options to cope with this issue: use advanced parallel hardware and parallelized data-mining algorithms or use sampling to reduce the size of the database to be mined.
- Sampling represents a trade-off between the choice of the “least number of samples that represent the entire database closely enough” for the question we are trying to answer.

- Based on how this choice is made, sampling methods are classified as being Static or Dynamic.
- Static Sampling is so called because it uses *static* statistical tests to decide if a sample closely represents the entire data set.
- This approach suffers from the obvious drawback that it does not take into account the data mining tool, that is going to be used on the obtained samples.
- Dynamic Sampling as opposed to Static Sampling uses the knowledge of the specific data mining tool in its decision as to the sufficient size of the sample.
- While Static Sampling often uses statistical hypothesis tests, the authors propose a “Probably Close Enough” criterion that takes the data mining tool into consideration for the trade-off decision.

Static Sampling Criteria

- Static sampling needs to decide on the size of the sample that is sufficiently similar to the parent database.
- A static statistical test hypothesis could be that each field of the sample comes from the same distribution as the original database.
- The hypothesis tests that are developed based on statistical theory are usually designed to minimize the probability of falsely claiming that 2 distributions are different.

- Ex. In a 95% level hypothesis test, assuming the 2 samples come from the same distribution, there is a 5% chance that the test will incorrectly reject the null hypothesis that the distributions are same.
- A 5% null hypothesis test, which are conservative about claiming 2 distributions are same was used in the test by the authors.
- Given a sample, static sampling runs the hypothesis test on each of its fields. If it accepts all of the null hypothesis then it claims that the sample does come from the same distribution as the original database, and it reports the current sample as sufficient.

Shortcomings of the Static Sampling Model

- When running several hypothesis tests, the probability that at least one of the hypothesis is wrong increases with the number of tests.
- It is unclear how the setting of the confidence levels will effect the sample size and performance of the sampling algorithm.
- The fundamental drawback is that when we want to know “if a sample is good enough” the answer is heavily dependent on the “what we are going to do with the sample”.

Dynamic Sampling

- A “Probably Close Enough(PCE)” criterion is used to address the trade-off decision in the case of Dynamic Sampling
- The PCE Criterion is a way of evaluating a sampling strategy. The key is that the sampling decision should occur in the context of the data mining algorithm we plan to use.
- The idea is to think about taking a sample that is “probably good enough” meaning that there is only a small chance that the mining algorithm could do better by using the entire database instead.
- A naïve Bayesian classifier was chosen to test the PCE framework.

- We would like the smallest sample size n such that

$$\Pr(\text{acc}(\mathbf{N}) - \text{acc}(\mathbf{n}) > \varepsilon) \leq \delta$$

where $\text{acc}(n)$ refers to the accuracy of the mining algorithm after seeing a sample of size n , $\text{acc}(\mathbf{N})$ refers to the accuracy after seeing all records in the database, ε is a parameter describing what “close enough” means, and δ is a parameter describing what “probably” means.

Sampling through Cross Validation

- In using the PCE criterion, we examine samples of increasing size n , adding a constant number of records to our sample repeatedly until we believe that the PCE condition is satisfied.
- First $\text{acc}(n_i)$ is estimated by using the leave-one out cross-validation on the sample.
- For a first attempt at the estimation of $\text{acc}(N)$, assume that whenever

$$\text{acc}(n_{i+1}) \leq \text{acc}(n_i),$$

the derivative of accuracy with respect to the training set size has become non-positive and will remain so for increasing sample sizes.

Thus

$$\mathbf{acc(N) \leq acc(n_i)}$$

and we should accept the sample of size n .

- The use of this method for putting a bound on $acc(N)$ was found to be sensitive to the variance in the estimates of $acc(n_i)$ and also was found to stop too soon.
- On the average, the accuracy was reduced about 2% from the accuracy on the full database while the sample size was less than 20% of the size of the original database.

Extrapolation of Learning Curves

- A better and more general method of estimation of $\text{acc}(N)$ uses all available data on the performance of the mining algorithm on varying-sized training sets, and fits this data to a parametric learning curve which is a function of the algorithm's accuracy w.r.t the size of the training sample.
- This curve can then be extrapolated to predict the accuracy of the mining algorithm on the full database.

- To fit the learning curve, we need an estimate of $\text{acc}(n)$. When we consider a sample of size n , we take K more records from the large database and classify them and measure the resulting accuracy. We use the history of sample sizes (of earlier smaller samples) and the measured accuracies to estimate and extrapolate the learning curve.

- A power law curve of the form

$$\text{acc}(n) = a - bn^{-\alpha}$$

was found to be a good fit to the learning curve through theoretic studies.

- The parameters a , b , α are fit to the observed accuracies using a simple function optimization method.
- This learning curve is used to estimate the accuracy of the data mining algorithm after seeing all N cases.

- This value of $\text{acc}(N)$ is compared with the accuracy on the current sample of size n and if the difference is not greater than ϵ , we accept the sample as being representative.
- If the difference is greater than ϵ , we reject the sample and add K additional records (sampled previously to get an estimate of the model) to the sample updating the model built by the algorithm.

Experiments ELC Vs Static Sampling

- 5-fold cross validation was used to test the accuracy.
- For each training step, the database was sampled using either no sampling (all records), static sampling with a 5% confidence level test and dynamic sampling with a $\text{acc}(N) - \text{acc}(n) < 2\%$.
- Both sampling algorithms were initialized with a sample size of 100 and incremented by 100 sample if it was ruled insufficient.

Table of Results

Table 2: Sample size (n) and accuracy for 25 runs.

Data set	Naive	Static		Dynamic	
	Acc.	Acc.	n	Acc.	n
Breast Cancer	95.9	95.9	300	95.9	300
Credit Card	77.7	77.0	500	77.2	1180
German	72.7	63.8	540	71.8	2180
Glass2	61.9	60.0	100	61.9	720
Heart Disease	85.1	83.2	180	85.1	900
Hepatitis	83.8	83.2	100	83.8	540
Horse Colic	76.6	76.1	240	76.6	640
Iris	96.0	96.0	100	96.0	560
Lymphography	69.1	67.1	100	68.5	600
Pima Diabetes	75.3	75.7	420	75.5	1080
Tic-tac-toe	69.7	69.2	620	71.1	620

Results

- 1) By using the PCE criterion with $\varepsilon = 0.02$, in no case was the accuracy of the entire database more than 0.9% higher than the extrapolated sample accuracy.
- 2) Static Sampling while providing smaller samples, did worse at matching the accuracy of the on the entire database, in 2 cases the accuracy was 1.9% worse than the accuracy on the entire database, and on one domain the accuracy was nearly 10% lower.

Conclusions

- Data mining process is a collaboration between a customer and a data analyst.
- Decisions about how large a sample is to be used must be made rationally.
- The authors propose a dynamic sampling algorithm that offers parameters that relate directly to the performance of the resulting model, rather than to a statistical criterion which is related in some unknown way to the desired performance.

Cribs and Gripes

- The paper is a very high level presentation of information. It assumes prior knowledge of statistical theory and machine learning procedures.
- The actual algorithm being proposed is very poorly described although the theoretical basis is discussed. It takes some work to piece together the sequence in which things are done.