THE EFFECT OF SAMPLING TECHNIQUES TO ACCURACY ESTIMATION

Kittisak Kerdprasop, Nittaya Kerdprasop, Pongden Punpakdeewong, and

Petchpirin Doungsuwan

School of Computer Engineering
Suranaree University of Technology
111 Muang District
Nakorn Ratchasima 30000
Phone (044) 224352; Fax (044) 224165
Email: kerdpras@ccs.sut.ac.th, nittaya@ccs.sut.ac.th, then007@yahoo.com, petch_winny@yahoo.com

ABSTRACT

Knowledge discovery is the process of extracting useful and previously unknown information from the very large data set. Among many discovering methods, decision rules extracting is one of the most extensively studied techniques. But extracting rules from a large database is computationally inefficient. Using a sample from the database can speed up the data mining process, but this is only acceptable if it does not reduce the quality of the induced rules. We thus investigate the criteria to decide whether a sample is sufficiently similar to the original database. We observe the accuracy of the induced rules extracted from training samples of decreasing sizes and use these results to determine when a sample is sufficiently small, yet maintain the acceptable accuracy rate. We evaluate random and systematic sampling methods on data from the UCI repository.

1. INTRODUCTION

Data mining (also known as knowledge discovery in databases, or KDD) is the process of applying specific learning algorithm to extract interesting and useful knowledge from data [2]. Typical data mining applications extract knowledge from databases ranging from small to moderate in size. When a data set is very large, mining process may take a very long time. Moreover, some mining algorithms may not be scalable on huge amounts of data. To handle large data sets, data reduction is one important step prior to applying the mining algorithms.

Data reduction can be achieved by reducing the number of cases and/or reducing dimensions of those cases. Our study focuses on case (or instance) reduction via the technique of sampling. Mining on reduced data set is obviously more efficient than on the original data set. On the contrary, if the sample is too small, some useful knowledge may be overlooked. Our paper addresses the question of sufficient sample size as well as the improved mining time. The rest of the paper is organized as follows. The next section describes various sampling methods. Section 3 explains the methodology of our study including experimental setup and the data sets. Section 4 discusses the results. The conclusion is presented in Section 5.

2. SAMPLING METHODS

Sampling is used as a data reduction technique because it allows a large data set to be represented by a much smaller subset of the data. Basic methods of sampling

commonly used are random sampling, systematic sampling, and stratified sampling [4].

Suppose that a large data set contains N instances. Random sampling selects n instances ($n < N$) at a random choice. The probability of drawing any instance in the data set is $1/N$, that is, all instances are equally likely. This is the case of random sampling without replacement. If the sampling is done with replacement, an instance has a chance to be drawn more than one times.

The systematic sampling method draws n instances from the data set by their fixed stepping positions. This sampling method draws the first instance at a random position. Then iteratively draws subsequent instance at the next k position, when k is a stepping size.

Stratified sampling method first divides the data set into mutually disjoint subsets called strata. Then draws samples from each stratum independently by applying the simple random sampling technique. The three sampling methods are in illustrated in Figure 2.1.

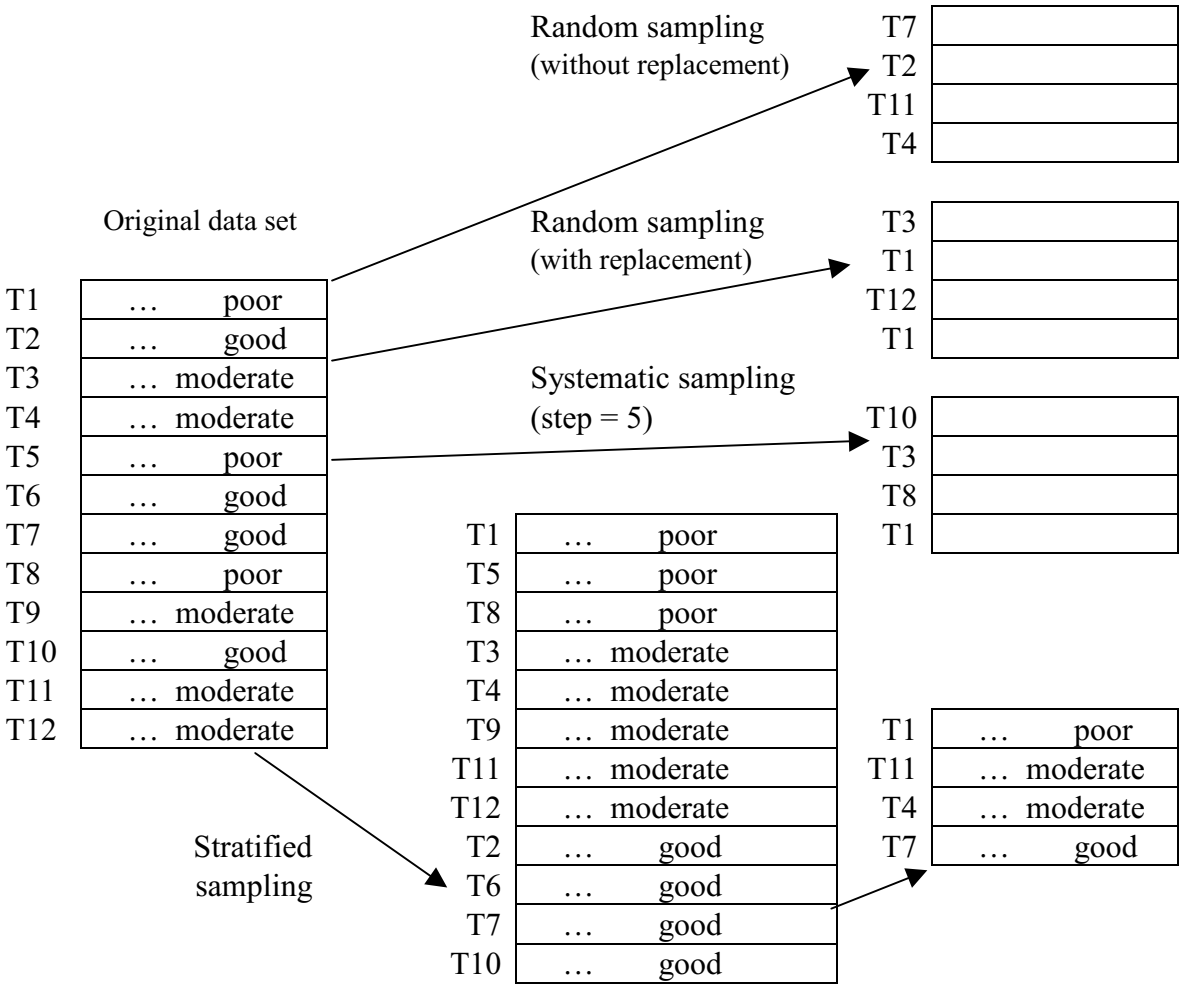

Figure 2.1 Different sampling methods to draw 4 samples

3. METHODOLOGY

3.1 Experimental Setup

In order to conduct an experiment to investigate the sufficient size of a sample obtained from different sampling methods, we use OneR as a learning algorithm to induce decision rules. OneR algorithm [3] induces decision rules based on the value of a single attribute. It is shown that we can get reasonably accurate decision rules by simply looking at one attribute, as opposed to a more sophisticated top-down decision-tree induction algorithms such as C4.5 [6]. The average accuracy of OneR for the data sets tested by Holte [3] is just 5.7% lower than that of C4.5.

We choose two data sets from the UCI Repository [1]. The two data sets represent the variety of data characteristics, that is, numeric data, nominal data, and data with missing values. Each data set is sampled using two different sampling methods: random sampling (without replacement) and systematic sampling.

For each sampling method, a data set is drawn for five different sample sizes: 80%, 70%, 60%, 50%, and 40% of the original data set. Then runs the learning algorithm on each sample. The learning algorithm is also run on the original data set to observe the accuracy and the learning time. These two criteria will be used as a benchmark to compare against those obtained from the various samples.

The experiments are performed on the WEKA (Waikato Environment for Knowledge Analysis) system [7]. WEKA system is an open-source Java-based machine learning environment that provides tools and algorithms to be used as a data- mining workbench.

3.2 Methods for Accuracy Estimation

Accuracy estimation refers to the process of approximating the future performance of a set of decision rules induced by a learning algorithm. This process helps to evaluate how accurately the induced rules will predict on the future data. The common accuracy estimation methods used extensively are [5] holdout, cross-validation, leave-one-out cross-validation, stratified cross-validation, and bootstrap. In our experiments, we estimate the accuracy of the induced rules using the holdout and stratified 10-fold cross-validation methods.

The holdout method partitions the original data set into two mutually disjoint sets: a training set and a test set (or holdout set). Typically, two thirds of the data ($2/3 \approx 66.6\%$) are allocated to the training set, and the remaining ($1/3 \approx 33.3\%$) is allocated to the test set. The training set is used to train the learning algorithm, and the induced decision rules are tested on the test set. Since only 33.3% of the original data are used for estimating accuracy, this method is not a good estimator for a small data set.

Ten-fold cross-validation method is a variation of the holdout method in which the method is repeated 10 times. The data set is randomly split into ten mutually disjoint subsets, called the folds. The ten folds are approximately equal in size. To induce decision rules, nine folds are used to train the learning algorithm, the remaining one fold will be used as a test set. The process is repeated 10 times with a different set of test data at each iteration. The overall accuracy estimation is the average of the accuracy obtained from each iteration.

Stratified 10-fold cross-validation is a cross-validation technique in which the original data is stratified before it is partitioned into ten folds. This is to guarantee the equal distribution of data in each class. This method is a good estimator for a small data set [5].

In order to observe the predicting accuracy of different sampling sizes, we employ both the holdout and the stratified

10-fold cross-validation methods since we range the sampling size from 100% (i.e., no sampling at all) to as small as 40% of the original data size.

4. RESULTS AND DISCUSSION

For each data set, we first apply sampling methods to generate the samples of different sizes. Each sample is then applied to the OneR learning algorithm to induce the decision rules. The predicting accuracy of these rules is tested with the two estimating methods: holdout and stratified 10-fold cross-validation. Tables 4.1 and 4.2 show the results of the accuracy estimation for the data sets using different sampling methods at various sizes. The columns 'Correct Rules' indicate the sampling sizes that generate the same set of rules as the original data (i.e., no sampling data set).

Table 4.1 Accuracy estimates for the Iris data set (numeric data)

Sampling Method	Sampling Size	Number of Instances	Accuracy (Holdout)	Correct Rules	Accuracy (stratified 10-fold CV)
No sampling	100%	150	98.0392%		92.6667%
Random Sampling	80%	120	87.8049%		92.5%
	70%	106	89.1892%	✓	94.3396%
	60%	90	93.5484%		96.6667%
	50%	75	92.3077%	✓	93.3333%
	40%	60	100%		100%
Systematic Sampling	80%	120	92.6829%		92.5%
	70%	105	97.2222%	✓	98.0952%
	60%	90	90.3226%		96.6667%
	50%	74	100%	✓	97.2973%
	40%	60	100%	✓	96.6667%

Table 4.2 Accuracy estimates for the Primary-Tumor data set (nominal data with missing values)

Sampling Method	Sampling Size	Number of Instances	Accuracy (Holdout)	Correct Rules	Accuracy (stratified 10-fold CV)
No sampling	100%	339	28.4483%		27.4336%
Random Sampling	80%	264	26.6667%	✓	28.0303%
	70%	231	26.5823%		24.6753%
	60%	199	32.3529%	✓	30.6533%
	50%	165	28.0702%		30.303%
	40%	132	35.5556%		32.5758%
Systematic Sampling	80%	264	26.6667%	✓	30.6818%
	70%	231	22.7848%	✓	29.4372%
	60%	198	27.9412%	✓	28.7879%
	50%	165	31.5789%		30.303%
	40%	132	31.1111%		26.5152%

For each data set varying in the characteristics, the results indicate that:

(1) For the small data set (Iris data), stratified 10-fold cross-validation estimating method on sampled data predicts with a higher accuracy than predicting with the original data.

(2) Sampling sizes of 50-70% produce the same set of decision rules as the original data, but the accuracy rate is higher.

(3) For a larger data set (Primary-Tumor data) with missing values, sampling sizes of 60-80% generate the same results as the original data, but with a higher accuracy.

(4) On average, random sampling generates slightly more accurate samples than the systematic sampling method on a larger data set.

(5) The time to induce the decision rules is 0 second on every sample. So, we cannot conclude that the smaller samples help decreasing the learning time.

5. SUMMARY

We review common sampling methods that are naturally used as a technique to reduce the data size for the purpose of improving learning time in the data mining process. We design the experiments to vary the sampling sizes in order to observe the smallest sampling size that yields the learning result as accurate as the original data. The accuracy is estimated with two different methods: the holdout and the stratified 10-fold cross-validation.

Our results indicate that random sampling of the size approximately 60-80% of the original data produces the same result as the original data with a high accuracy. However, to draw the exact conclusion requires a further investigation on a larger data set as well as with two or more learning algorithms.

REFERENCES

[1] C. L. Blake and C. J. Merz: "UCI Repository of machine learning databases," University of California, Irvine, Department of Information and Computer Science, 1998. [http://www.ics.uci.edu/~mlearn/MLRepository.html].

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth: "From data mining to knowledge discovery in databases," AI Magazine, pp.37-54, 1996.

[3] R.C. Holt: "Very simple classification rules perform well on most commonly used datasets," Machine Learning, Vol. 11, pp.63-90, 1993.

[4] K. Josien, G. Wang, T.W. Liao, E. Triantaphyllou, and M.C. Liu: "An evaluation of sampling methods for data mining with fuzzy c-means," In Dan Braha, editor, Data Mining for Design and Manufacturing, Chapter 15, pp.351-365, Kluwer Academic, 2001.

[5] R. Kohavi: "A study of cross-validation and bootstrap for accuracy estimation and model selection," Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), pp.1137-1143, 1995.

[6] J. R. Quinlan: "C4.5: Programs for Machine Learning," Morgan Kaufmann, 1993.

[7] WEKA (Waikato Environment for Knowledge Analysis), University of Waikato, Department of Computer Science, New Zealand. [http://www.cs.waikato.ac.nz/~ml].