The Nucleus **50**, No. 3 (2013) 219-228

The Nucleus
A Quarterly Scientific Journal of Pakistan Atomic Energy Commission
NCLEAM, ISSN 0029-5698

OPTIMAL SAMPLING STRATEGY FOR DATA MINING

*A. GHAFFAR, M. SHAHBAZ and W. MAHMOOD[1]

Department of Computer Science and Engineering, University of Engineering and Technology, Lahore, Pakistan

[1]Al-Khwarzmi Institute of Computer Sciences, University of Engineering and Technology, Lahore, Pakistan

(Received May15, 2013 and accepted in revised form August 29, 2013)

Latest technology like Internet, corporate intranets, data warehouses, ERP's, satellites, digital sensors, embedded systems, mobiles networks all are generating such a massive amount of data that it is getting very difficult to analyze and understand all these data, even using data mining tools. Huge datasets are becoming a difficult challenge for classification algorithms.

With increasing amounts of data, data mining algorithms are getting slower and analysis is getting less interactive. Sampling can be a solution. Using a fraction of computing resources, Sampling can often provide same level of accuracy. The process of sampling requires much care because there are many factors involved in the determination of correct sample size. The approach proposed in this paper tries to find a solution to this problem. Based on a statistical formula, after setting some parameters, it returns a sample size called *"sufficient sample size"*, which is then selected through probability sampling. Results indicate the usefulness of this technique in coping with the problem of huge datasets.

Keywords: Probability sampling, Classification, Sample size, Decision tree, Data mining

1. Introduction

Data mining is generally defined as "The nontrivial extraction of implicit, previously unknown and potentially useful information from data" [1] or "the science of extracting useful information from large data sets or databases." [2, 3]. Traditionally, statisticians, operation researchers & business analyst's job routinely involve some kind of manual extraction of useful knowledge from data(data analysis). But the increasing amounts of data are making their job more and more difficult. In addition to that the rapid increase in technology has made collection of huge amounts of data very easier. Sensors, computers, mobiles, internet, corporate databases all are generating huge amounts of data. It is not possible to analyze all this data humanly, so there should be some automated process.

Common data mining techniques include association rule mining [4], classification [5] and clustering [6]. Classification technique will be the main focus in this study.

Classification is supervised learning technique which is used to learn concepts (class labels) from given data, these concepts are used to label the data, whose class label is unknown [7]. A survey and experimental comparison of classification algorithms is given [5].

This paper is organized into 7 sections. Section 2 describes the problem statement. In section 3, an overview of classification and sampling is given. Section 4 is literature survey of various sampling methods and their application in classification. In section 5, a methodology will be proposed to manage with the problem of huge datasets and explain why proposed approach is better as compared to other approaches. In section 6, proposed methodology will be validated by presenting case study. In section 7, paper will be concluded.

2. Problem Statement

The volume of data grows too fast for hardware to keep up. With the current advancement in computing technology, we witness a rapid increase in the involvement of computer in every field of life. Gradually computers are taking over every field and resultantly the amount of data being generated is increasing exponentially. Internet corporate intranets, data warehouses, ERP's, satellites, digital sensors, embedded systems,mobiles all are generating such a massive amount of data that it is getting difficult to analyze and understand all these data, even using data mining tools. With increasing amounts of data, data mining algorithms are getting slower and analysis is getting less

* Corresponding author : abidghaffar@gmail.com

interactive. Large data sets can cause problem in two dimensions.

- Processing Delays.
- Classification Model Complexity.

The first thing is processing delays. For example in the context of decision tree approach, as evident from its algorithm[8], on root node, all the data instances have to keep in RAM. RAM is seldom sufficient, when big data sets are involved, so virtual memory has to be used. Virtual memory by its very nature is much slower than RAM, and this introduces delays in processing. At each tree subsequent node this process is repeated. So anyone can imagine the amount of processing. This all increases I/O & processing exponentially and creates a major bottleneck.

The other problem is classification model complexity. The DT is constructed in recursive way, which results in very complex tree which often over fits the data [9]. Most recent algorithms of DT e.g. C4.5 normally apply a step called as 'pruning [10] to simplify the tree. However, even the simplified tree does not solve the entire problem. It has been reported that the simplified tree size is roughly proportional to training data [11]. This makes it difficult to interpret and understand, hence is not of much use.

Briefly

- The amount of data is increasing exponentially.
- Data mining Algorithms need to see these data instances more than once.
- Data Mining Algorithms have super-linear time complexity in terms of training data instances. Exact results take impractical time and it reduces interactivity.
- The resultant model gets very complicated, when it comes to very huge data sets.

3. Classification and Sampling

3.1. Classification

Classification is supervised learning technique which is used to learn concepts (class labels) from given data, these concepts are used to label the data, whose class label is unknown [6].

Formally, the classification problem can be viewed as a function Φ (called a classifier) that maps an instance vector x = {a1, a2,, ..., ar} to the class label y € {L1, L2,..., LC} of the instance. Once we find such a model y = Φ(x), we could predict class label of new coming instances.

Note that this description is not that strict in mathematics, because in the context of classification [12].

- The exact solution of the function may not exist or be extremely hard to find, thus it is often solved approximately
- The input vector could contain non-continuous values, e.g., nominal (or categorical) values.

There are many approaches to classification for example Decision Tree, Neural Network, RBF, Gaussian Mixture Model, Gaussian, Support vector Machine, Discriminate Analysis, Bayesian network and k-nearest neighbor [13]. In this paper, I will mainly use Decision tree approach.

Decision tree is a tree-like tree structure whose internal node denotes a test on an attribute, each branch represents an outcome of the test and leaf nodes represent class labels or class distribution. The generation of DT is termed as Tree induction [14].

DT approach is used most in classification because it according to [15] provides easy to understand result which is both accurate and efficient.

3.2. Sampling

Sampling is the process selecting samples from the population. In simple words it is a procedure by which we generalize the results of part of a population to infer about the population.

- *Population* isset of observations about the subject under consideration. For convenience it could be thought as the Universal Set. In terms of classification, entire data set could be regarded as the population.
- A sample is a subset of population. Every sample is characterized by its size. No of observations present in the sample is called its sample size.
- A Representative sample has all the important characteristics of the population from which it is drawn. The best sampling size is often a tradeoff between what is desirable and what is practically feasible.
- Sampling Unit is the thing, which is sampled, for example, a person, a clinical episode, or a health facility. In classification, this could be termed as a data instance.
- Sampling method is a procedure by which sampling units are selected in population.

There are two main categories of sampling methods.

A survey about various data reduction techniques can be found in [16]. A study report about comparison of various constructing summary data techniques can be found in [12]. The contemporary approaches to instance selection methods can be found in [8].

4. Related Work

According to Gay and Diehl [17], for descriptive research the sample should be at least 10%-20% of population based on population size. Sampling has been used to cope with the problem of larger datasets in the following ways:

4.1. Sampling as a Wrapper

In this manner, the sampling strategy is embedded in the data-mining algorithm. Used in this manner, the sampling algorithm is a constituent part of the classification process. Some examples of algorithms that use sampling in wrapper manner are windowing [9], BOAT [18] and Peepholing. Sampling used as a wrapper, helps to resolve the problem of efficiency, but fails to build simple & interpretable models [11], that are easy to understand. It could be attributed to the fact that these classification algorithms, although initially select a sample from the whole data set, they gradually use the whole training data set over several iterations, in each of which iterations, a sample whole of the data is used.

In 'sampling as a wrapper' strategy, sampling is tightly coupled with the algorithm. Although the algorithm may solve some issues of classification, another algorithm or application cannot utilize the features of algorithm of 'sampling as a wrapper'.

So in short, sampling used in wrapper manner could not be used as a general solution[19].

4.2. Sampling as a Filter

In this manner, the sampling strategy is separated from data mining algorithm. Used in this manner, the sampling algorithm performs first and the sampled data is then delivered to data mining algorithm. Some examples of algorithms that use sampling in filter manner are

4.2.1. Dynamic Arithmetic Sampling

The dynamic arithmetic sampling was proposed by John and Langley [7]. It works on the theory of learning curve. Starting from a small sample size, it works in an incremental; iteratively it draws larger samples to learn, until a PCE (Probably Close Enough) criterion is met. The PCE is defined as

$$Pr(acc(N) — acc(n) > e) < \delta$$

Where acc(n) is result obtained from sample of size n and acc(N) is the population result, e is the loss of tolerable accuracy, and δ is error limit. The PCE is evaluated by fitting the learning curve with the power law. One obvious problem is that dynamic sampling can only work for incremental learning classifier like naïve Bayesian classifiers. Its performance as compared to other techniques is poor [20]. The other problem is that it heavily relies on the concept of learning curve. Basically it assumes that gradually the learning curve will become flat. However, some experiments deny that. Another issues that they tested their results on inflated datasets, which can cause concern. Even if everything goes well, there has been witnessed much wastage of time & space due to the successive iterations by Provost et al. [20]. Also some modification had been suggested by Provost et al. [20] to decrease these number of iterations.

4.2.2. Dynamic Geometric Sampling

The dynamic geometric sampling was first proposed by Provost et al. [20] for improvements in the dynamic arithmetic sampling algorithm. To achieve the largest possible gain in efficiency, the sample size is suggested to be increased geometrically, i.e., $nk+1 = a^* nk$, where a is a constant. Basically it uses the idea similar to the dynamic arithmetic sampling, but it does not require the algorithm to be incremental. It starts learning with a small sample, and progressively draws larger sample until classification model accuracy no longer improves. The improvement in model accuracy is evaluated by fitting the learning curve. It works in the following way.

- It obtains samples of 100,200,300,400,500 instances and building model upon them
- Estimating a power function based learning curve based on results of those models
- Selecting the next sample size to be the size required to achieve the accuracy criteria according to learning curve.

However, there are some issues with this technique. First is that it heavily relies on the concept of learning curve. Basically it assumes that gradually the learning curve will become flat. However, some experiments deny that [21]. It also

does not provide an efficient procedure to determine the starting sample size.

4.2.3. Static Sampling

In this approach a sample is drawn and its similarity is compared with the population (mother data). The basic approach is to test the hypotheses that each filed of sampled data is from the original population.

For categorical fields, an X2 hypothesis test is used to test the hypothesis that the sample & mother data has same distribution. For numeric fields, a sample is drawn such that the sample and mother data has same mean. However, this generalization could forfeit confidence about the dependencies the sample should accede to. Give a sample, static sampling test the suitable hypothesis. If all the null hypothesizes are successful, then the sample is accepted as valid and current sample size is said to be sufficient.

There are several issues with static sampling. First, hypothesis needs some presumed theories, say theories about distributions, which we usually do not know beforehand. Second, hypothesis tests have difficulties in controlling Type I Error and Type II Error [22, 23]. Third, how to set an exact null hypothesis is also hard. Fourth, when running several hypotheses, the probability of having at least one wrong test increases with increase in number of attributes. In short, we need an alternative to hypothesis testing for measuring sample quality.

5. Proposed Methodology

There is a need to develop such a technique, which could significantly speed up the classification process while having negligible effect on classification accuracy. One possible solution was proposed as using a scaled down sample of the whole data, which is small enough to be efficient and accurate enough to give very good approximation to the results. This should be done with extreme car because there is a tradeoff between sample size and accuracy. In general, the more data, the better results. But more data also means more processing. A sample size begins to grow; we begin to lose the advantages of sampling. There are many factors involved in the determination of correct sample size. So the decision about correct sample size requires much care and is very crucial.

Finding the correct sample size has long been a topic of research in statistics and considerable work has been done on the topic. But the work by Krejcie et al. [24] and Cochran et al. [22] are considered benchmark in this regard.

Unfortunately, on the other hand, there is not much research available on the topic of sufficient sample size in the context of classification.

The proposed approach uses probability sampling. Based on a statistical formula, after setting some parameters like sample quality and risk factor it returns a sample size, we will use “random sampling without replacement” to get “sufficient sized sample” provided by formula mentioned [24]. There are two parts of proposed approach. One is the use of probability sampling, and other is the use of sampling formula proposed by Krejcie et al. [24]. The algorithm for this approach is as under:-

1. Select data set to be analyzed.
2. Select desired confidence level of end result.
3. Select desired precision level.
4. Get sufficient sample size using proposed formula.
5. Retrieve the probability sample from the given population.
6. Return the sample.

To calculate a sufficient sample size, this approach provides great flexibility. The confidence level & precision level can be chosen keeping in view the requirements. If highly accurate results are required, confidence level of 99.9 % & precision level .01 % can be chosen. On the other hand if he just wants to get an overview of results, he can for example, select 95% confidence level and 3 % precision level. So using levels of precision & confidence, this approach provides wide choices and serves as a general solution to the problem of huge data sets.

5.1. Background Theory

Central limit theorem[25,26] is the second most important theorem of statistics that is central to the use of statistics. According to this theorem for any population with mean μ and standard deviation σ, the distribution of sample means for sample size n will approach a normal distribution with a mean of μ and a standard deviation of σ as n approaches infinity [27].

More formally if X1, X2, X3, ... Xn represent a series of n independent random variables not having infinite expectation μ and variance $\sigma 2 > 0$. Then according to central limit theorem, as the sample size n increases, the distribution of the

sample average of these random variables approaches the normal distribution with a mean μ and variance σ2 / n irrespective of the shape of the original distribution[28].

Let the summation of n random variables be Sn,

$$Sn = X1 + ... + Xn$$

Then, defining a new random variable

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{\mu}}$$

Distribution of Zn becomes the standard normal distribution N(0,1) as n approaches ∞ this could be written as

$$n\sqrt{n(\bar{X}_n - \mu)} \xrightarrow{D} N(0, \quad \sigma^2),$$

Where

$$\bar{X}_n = \frac{S_n}{n} = (X_1 + ... + X_n)/n$$

Is the sample mean.

5.1.1. Explanation

The average or mean summarize something known as central tendency. This is an easy way to tell what is normal, typical or expected in a population. For example, if average height of boys in a town is 5.10 ft. then any boy with unknown height could be expected to have height of approximately 5.10 ft.

The other measure which let us better approximate the any typical or normal in a population is variability or spread. This is formally called variance. Variance lets us know about the variability of the population. The square root of variance called standard deviation, tells us that how much each observation deviates from the average. Keeping in view the previous example. , if the boys have an average height of 5.10 ft. with a standard deviation of 3 inches, we could be very sure that any boy with unknown height would have a height range of 5.7 to 6.1. In fact, as statistics "66-95-99 rule" says, 65 % of the boy's height would fall between 5.7to 6.1 ft. 95% of the boys height would be between 5.4 to 6.4 ft and 99% of boy's height would fall in the range 5.1 to 6.7 ft.

Average and standard deviation both are parameter measures (the measure which we get by manipulating all the instances of a population is called parameter). While average & standard deviation together gives us very precise measures to describe a population, then calculation involves manipulation of every instance of a population. With millions & billions of instances, this could be very problematic. As everyone can estimate that vast amount of resources are required to process every instance of a huge population.

Due to the properties of central limit theorem, sampling distribution will act more and more like normal distribution as the sample size is increased, even when the population itself is not normally distributed [26].

5.2. Sufficient Sample Size

To obtain sufficient sample size, following parameters are required [1].

- *Population Size:* The size of population.
- *The Confidence Level:*The degree of confidence required in sample.
- *The Precision Level:* precision level required. This also controls the natural error of sampling.
- *Variability Degree*: The amount of wilderness or variability of population. In this paper, a constant degree of 50 % is assumed.

5.3. Krejcie's Formula

Using following statistical formula, we will determine a "sufficient sample size".

$$Sample\ Size = \frac{X^2(NP(1-P))}{d^2(N-1) + X^2P(1-P)}$$

Where

X = Confidence Level (Table value of chi square)

N = Population size

P = Population proportion (Assumed to be 0.5)

d = The Precision level.

5.4. Benefits of Proposed Methodology

This formula will help classification algorithms in following ways.

- The proposed approach uses statistical formula based on Krejcie et al. [24] research paper. This formula is being used successfully in numerous other research areas. According to [22], more than 450 other research papers have cited this article. Its use is also very common among scientists across a broader spectrum. In addition to it, this formula also gives us an estimate of error present in our sample, a dimension that is not available in

other approaches. So our approach relies upon a solid statistical background.

- Proposed approach uses parameters for sample size estimation. The end user is able to select the sample quality & risk factor according to his own needs. So he is in control. He can decide, what balance among learning efficiency, learning accuracy and the model complexity is desired. If, for example he just wants to get a rough estimate of end result, he could use a combination of low quality & moderate risk factor. On the other hand, if he wants to get precise results, he can opt for high quality & low risk sample. This is all matter of his preferences. So according to his requirements, he is able to make a decision.
- In terms of accuracy, although probability sampling is very simple strategy, it is best strategy compared to other instance selection strategies, as stated by Syed et al. [29] in their study,. By selecting only part of total data, sampling can, to some extent, decrease the amount of noise present in data. By reducing the amount of noise, classification accuracy could be improved. In addition to it sampled data also resolves the issues of 'overfitting' of model
- In terms of efficiency, any classification algorithm using the proposed approach could excel in efficiency, as it now had to process only a small fraction of data, and there is huge gain in terms of efficiency.
- In terms of scalability, when any classification model which process massive data very slowly, now will be able to have efficiently handle the sample data and produce the competitive results hence becoming more scalable than before. This "sufficient sized sample" normally consists of only "1-20 %" of original data. The size of sample depends on many factors. In addition to parameters of formula, the % of sufficient sample size decrease as the population size increases, so the real saving is in case of larger data sets.
- According to Oates et al. [30], larger datasets results in complex models, without any increase in model accuracy. So in terms of comprehensibility, the resultant classification model from sampled data will be far more simple and understandable as compared to the classification model built using all the data.
- Probability sampling can be done more efficiently than other sampling or instances selection methods. Most other sampling methods presume some kind of assumption like priori knowledge about data or data mining algorithm. Some other methods relies on the theory of learning curve (for example Progressive Sampling), which results in iteratively selecting and reviewing whether the learned accuracy is increasing or not. Some other recursively use all the data. This will cause extra computation cost in selecting a sample & subsequent processing. This results in loss of efficiency. A Probability sample of n records from a data set with N records can be obtained in at most O(N) time. When using probability sampling with a reservoir [31], the expected time can be reduced to O(n(1 + log(N/n)).
- Probability sampling produces unbiased samples of the target data and is independent of neither the used data-mining algorithm, nor the data. Given a huge amount of dataset, it will return a very small subset of the original dataset, which could further be used for data mining. Almost all other statistical sampling methods or instance selection methods assume some knowledge of the data set, or have some preferences on the individual instances, or require some knowledge of the data mining algorithms, thus are either unable to produce unbiased sample, or are dependent on the used data mining algorithms.
- A comparison chart of proposed methodology & other approaches is presented below.

6. Evaluation

In this section first a dataset for evaluation. Our evaluation consists of following steps.

- Data Preprocessing
- Sampling
- Model Building
- Validation

Letters data set provided by UCI machine learning repository, is selected for benchmarking. It consists of 17 attributes including one class attribute and 20000 instances. More information about this dataset could be found [32].

6.1. Data Preprocessing

In data preprocessing step, we will divide our dataset into two parts i.e. Training dataset and Validation dataset. In Classification, this partition is

done in such manner that 70%-80% data is used as training data. Using this training data a classification model is built. The accuracy of this model is verified using the remaining 20%-30% data that is termed as Validation dataset.

So, in this step, the *letters* dataset was divided into training dataset having 16000 instances & validation dataset having 4000 instances.

6.2. Sampling

In this step, we will calculate a *Sufficient Sample Size* for our training data by keeping in view the required accuracy. Once we calculate the desired *Sufficient Sample Size*, we will draw this sample using "Random Sampling without replacement".

Using the sufficient sample size calculator, sufficient sample size is calculated. These sample size is calculated for confidence level of 95% and precision Level of 2.50 % as given below.

Sample Size Calculator	
Confidence Level	95%
Percision Level	2.50%
Population	16000
Sufficient Sample size	**3998**

Figure 6. Sufficient sample size calculator.

A sample size of 3998 was given. It was rounded to make it 4000. 4 samples of each 4000 instances were drawn out of Training dataset.

6.3. Model Building

In this step, using classification algorithm, we build classification model using our training set. For our purpose, we will first build our model using the whole training dataset, and then build the classification model using the *"Sufficient Sample"*, we draw in previous step. Both the models will be compared to realize the benefits of using sampling approach.

Classification models for whole training dataset & sample datasets were build-using XLStat[33], Ctree[34] and Tanagra[35]. Multiple software & multiple algorithms were checked so that our methodology could be tested across a broader spectrum.

6.4. Validation

Classification models for whole training dataset & sample datasets were build using XLStat, Ctree and Tanagra. The performance of classification model would be discussed using following parameters

- Classification Accuracy over training dataset.
- Classification accuracy over Validation dataset.
- Running time taken.
- Model Complexity.

The results of first three parameters, i.e. classification accuracy of training data, classification accuracy of validation data & running time taken are summarized in the following figure.

	Classification Accuracy		Running Time(seconds)	
	Training	Validation	XLStat	Ctree
Ist Sample(20 % Data)	51.55%	33.70%	8	61
Ist Sample(20 % Data)	49.15%	30.53%	8	61
Ist Sample(20 % Data)	50.95%	35.88%	8	61
Ist Sample(20 % Data)	51.00%	33.28%	8	61
Average of four samples	50.66%	33.34%	8	61
Training Data (100% Data)	51.37%	35.15%	60	668
Improvement	-0.71%	-1.81%	86.70%	90.90%

Figure 7. Results summary.

As it is clear from this figure, by giving up an accuracy of 0.71 % for training data & 1.35 % for validation data, one is able to achieve a gain of 86.7 % using XLStat & 90.90 % using Ctree. That's a huge improvement. By giving up very small quantities in terms of accuracy, one can achieve substantial gains.

Based on obtained results, the relationship of classification accuracy with the sampling data (as % of whole data) is depicted in the following graph.

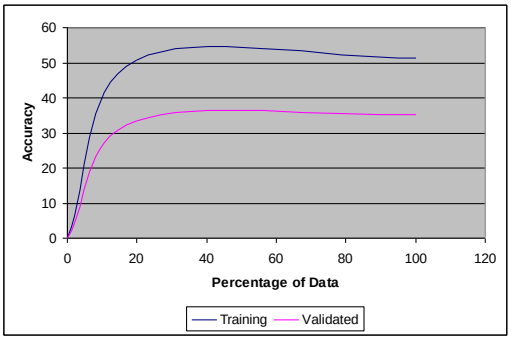

Figure 8. Sampling data & classification accuracy graph.

It is evident from graph that beyond 20% of data, the classification accuracies of training dataset and validation dataset does not improve much. This supports my claim regarding sampling. It is also worthnoting that both the curves closely resemble the famous “Learning Curve”. These resemblances also indirectly support our proposed methodology.

The relation of sampling size (as % of whole data) & running time (classification accuracy) is depicted in the following graph.

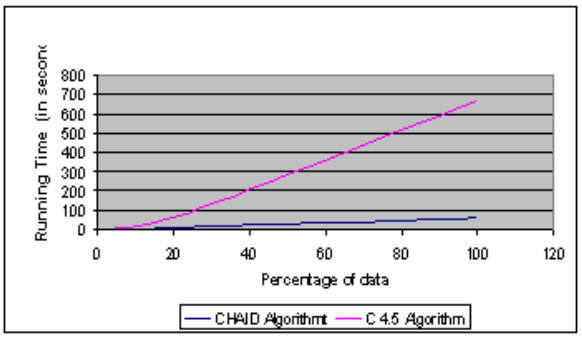

Figure 9. Sampling data & classification efficiency graph.

It is evident from graph that sampling could substantially reduce the time required for classification. In terms of model complexity, we can see the improvement in following diagram

	Number of nodes	Number of leaves
Training Data	1601	801
Sample Data	529	265
Difference	**1072**	**536**
Improvement	**66.96%**	**66.92%**

Figure 10. Classification model Complexity Comparison.

It’s evident from the figure that use of sample generated much smaller & interpretable tree. It reduced the complexity of model nearly 66 %. So classification model complexity could be greatly reduced by the use of sampling.

6.5 Evaluation Over Other Datasets

To broaden the scope of our approach, following three datasets were chosen from UCI repository[32].

- Adults
- Nursery
- Pendigits

The adult dataset was selected because, it contains a good mix of both thenumerical a well as categorical attributes. The nursery dataset only consists of categorical attributes and Pendigits only consists of numerical attributes. In addition to it all the three datasets are large enough to test for data mining. This makes a good testing challenge for our proposed strategy. These datasets were evaluated in the manner described in 4.3. The results of evaluation are as under.

Table 4.1. Description of 3 UCI data sets.

Data Set	Numerical Attributes	Categorical Attributes	Cases for Training	Cases for Testing
Adult	6	8	36000	12842
Nursery	0	8	10000	2960
Pendigits	16	0	8000	2992

For adult dataset, out of 36000 instance of training dataset were replaced with the 3719 instance of “sufficient sample”, which merely results in loss of 0.9%. In the same way, Nursery dataset consists of 10000 training instances which reduce to 2931 instances of “sufficient sample”, with loss of only 0.4% accuracy. We can also see that the same story repeats for Pendigits dataset.

Table 4.2. Summary results of 3 UCI datasets.

Data Set	Full Size & Accuracy	Sufficient sample size & Accuracy
Adult	36000, 85.8% (C4.5)	3719, 84.9% (C4.5)
Nursery	10000, 93.9% (C4.5)	2931, 92.9% (C4.5)
Pendigits	8000, 96.4% (C4.5)	2731, 96.3% (C4.5)

It’s evident from table that although “Sufficient sample size”, much smaller in size,it produces competitive results in terms of accuracy. The results clearly indicate the validness of our approach. By losing negligible in terms of accuracy, we gain substantial in terms of efficiency and model complexity.

7. Conclusion

This study uses a sufficient sample size for classification. When using sampling, one has to beware of sample size, model accuracy, and efficiency & model complexity. In the context of

applying probability sampling in classification, a major concern is to find the tradeoff among sample size, model accuracy and model complexity.

I have studied the approach of probability sampling alongwith the approach of finding sufficient sample size to solve the issues of classification. The first part of proposed approach is based on statistical theory and probability theory. Using statistical background, first part of my approach gives a sufficient sample size. By sufficient sample size, it means such a sample that gives a desired set of efficiency and accuracy. The main contribution of this thesis is

- Provide an overview of data mining and classification.
- Provide a survey on how sampling is used for classification.
- Suggest the use of sample size formula proposed by Krejcie et al. [24] alongwith probability sampling for sampling in the context of classification.

This approach used in this paper intends to serve as a general solution to the problem of large data sets. All the classification algorithms can benefit from it.

The approach adapted in this paper could benefit classification algorithms in following ways

- Increasing efficiency.
- Decreasing model complexity.
- Making it more scalable.
- Resolving overfitting issues.
- Increasing classification accuracy.

8. Future Work

Although the proposed methodology has been tested and validated for classification technique only, one can find its applications in other data mining techniques also. Effect of sampling on these techniques could possibly be an extension to this work.

References

[1] W. Frawley, G. Piatetsky-Shapiro and C. Matheus, Knowledge Discovery in Databases: An Overview, AI Magazine (1992) pp. 213–228.

[2] D. Hand, H. Mannila and P. Smyth, Principles of Data Mining. MIT Press, Cambridge, MA. ISBN 0-262-08290-X (2001).

[3] http://en.wikipedia.org/wiki/Statistical_ classification.

[4] H. Jochen, U. Guntzer and G. Nakhaeizadeh. Algorithms for Association Rule Mining – A General Survey and Comparison. SIGKDD Explorations, 2, No. 1 (2000) 58.

[5] S.K. Murthy. Automatic Construction of Decision Trees from Data: A Multi-disciplinary Survey. Data Mining and Knowledge Discovery **2** (1998) 345.

[6] J. Han and M. Kamber.Data Mining Concepts and Techniques, Morgan Kauf-mann Publishers (2000).

[7] A.K. Jain, M.N. Murty and P.J. Flynn, Data clustering: A review, ACM Computing Surveys **31**, No. 3 (1999) 264.

[8] H. Liu and H. Motoda. Instance Selection and Construction for Data Mining. Kluwer Academic Publishers (2001).

[9] B. Chandra, P. Paul Varghese. Information Sciences **179**, No. 8, (2009) 1059.

[10] R.J. Freund and W. J. Wilson. Statistical Methods. Academic Press, Inc., San Diego, CA, USA (1997).

[11] T. Lim, W. Loh and Y. Shih. A Comparison of Prediction Accuracy, Complexity and Training Time of Thirty-three Old and New Classification Algorithms, Machine Learning (1999).

[12] W. DuMouchel. Handbook of Massive Data Sets, Chapter Data Squashing: Constructing summary data sets, Kluwer Academic Publishers (2001) pp. 1-13.

[13] http://www.cs.sfu.ca/~han/bk/7class.ppt.

[14] http://scholar.google.com/scholar?q= Determining+Sample+Size+for+research+activities&hl=en&lr=&btnG=Search.

[15] http://www.cse.unsw.edu.au/~billw/cs9414/ notes/ml/06prop/id3/id3.html.

[16] D. Barbara, W. DuMouchel, C. Faloutsos, P.J. Haas, J.M. Hellerstein, Y. Ioan-nidis, H.V. Jagadish, T. Johnson, R. Ng, V. Poosala, K.A. Ross and K. Sevick. The New Jersey data reduction report. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering (1997).

[17] L.R. Gay and P.L. Diehl, Research Methods for Business and Management, New York, Macmillan (1992).

[18] J. Gehrke, V. Ganti, R. Ramakrishnan and W.Y. Loh. Boat- Optimistic Decision Tree

Construction, In Proceedings of SIGMOD'99 (1999).

[19] J. Catlett. Megainduction: A test flight. In Proceedings of the Eighth International Workshop on Machine Learning, Morgan Kaufmann (1991) pp. 596-599.

[20] F. Provost, D. Jensen and T. Oates. Efficient progressive sampling. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99), AAAI/MIT Press (1999) pp. 23-32.

[21] Baohua Gu, Random Sampling for Classification on Large Data Sets, MSc Thesis, National University of Singapore, (2002).

[22] http://en.wikipedia.org/wiki/Data_mining.

[23] G.H. John and P. Langley. Static versus dynamic sampling for data mining. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI / MIT Press (1996).

[24] N.A. Syed, H. Liu and K.K. Sung. A study of support vectors on model independent example selection. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99, 1999).

[25] http://www.intuitor.com/statistics/Central Lim. html

[26] http://mlr.cs.umass.edu/ml/datasets.html

[27] http://en.wikipedia.org/wiki/Central_limit_ theorem

[28] M. Marasinghe, W. Meeker, D. Cook and T.S. Shin, Using Graphics and Simulation to Teach Statistical Concepts, Paper presented at the Annual Meeting of the American Statistician Association, Toronto, Canada (August, 1994).

[29] Miaoulis, George and R. D. Michener, An Introduction to Sampling, Dubuque, Lowa: Kendall/Hunt Publishing Company (1976).

[30] T. Oates and D. Jensen. The Effects of Training Set Size on Decision Tree Com-plexity, In Proceedings of the Fourteenth International Conference on Machine Learning (1997).

[31] W.G. Cochran, Sampling techniques (3rd Ed.), New York: John Wiley & Sons (1977).

[32] http://mlr.cs.umass.edu/ml/datasets/Letter+R ecognition

[33] http://www.xlstat.com/Download.htm

[34] http://www.geocities.com/adotsaha/CTree/ CtreeinExcel.html

[35] http://eric.univ-lyon2.fr/~ricco/tanagra/ index. html.