A Comparative study of sample selection methods for classification

P E N Lutu*, A P Engelbrecht‡

*Department of Informatics, University of Pretoria, SOUTH AFRICA
‡Department of Computer Science, University of Pretoria, SOUTH AFRICA

RÉSUMÉ

L'échantillonnage pour le minage de large ensemble de données est important pour au moins deux raisons. Le traitement de grande quantité de données a pour résultat une augmentation de la complexité informatique. Le coût de cette complexité additionnelle pourrait être non justifiable. D'autre part, l'utilisation de petits échantillons a pour résultat des calculs rapides et efficaces pour les algorithmes de minage de données. Les méthodes de statistique pour obtenir des échantillons d'ensemble de donnés satisfaisants pour les problèmes de classification sont discutées dans ce papier. Des résultats sont présentés pour une étude empirique basée sur l'utilisation d'échantillonnage aléatoire séquentiel et l'évaluation d'échantillon utilisant le test d'hypothèse univariée et une mesure theoretique de l'information. Des comparaisons sont faites entre des estimations théoriques et empiriques.

MOTS-CLÉS: échantillonnage d'ensemble de données, analyse de données, apprentissage de machine, classification, information measures

ABSTRACT

Sampling of large datasets for data mining is important for at least two reasons. The processing of large amounts of data results in increased computational complexity. The cost of this additional complexity may not be justifiable. On the other hand, the use of small samples results in fast and efficient computation for data mining algorithms. Statistical methods for obtaining sufficient samples from datasets for classification problems are discussed in this paper. Results are presented for an empirical study based on the use of sequential random sampling and sample evaluation using univariate hypothesis testing and an information theoretic measure. Comparisons are made between theoretical and empirical estimates.

KEYWORDS: dataset sampling, data analysis, machine learning, classification, information measures

1 INTRODUCTION

Data mining activities based on machine learning algorithms such as artificial neural-networks [2], decision trees [20] and K-nearest neighbor [5], amongst others, require lengthy computation times due to the sophisticated functions used by these algorithms. For very large datasets, the amount of time required for the computations can quickly become infeasible.

There is plenty of evidence to suggest that it is not desirable to present very large datasets to inductive machine learning algorithms. Catlett [4], Kohavi [13] and Provost et al [19] have demonstrated that the learning curve for a very large dataset will normally flatten before the whole dataset has been used. Elomaa and Kaariainen [7] have however cautioned that not all domains have the typical well-behaved learning curve which rises steeply and then gradually flattens. For artificial neural networks, Engelbrecht [8] has demonstrated that even for small datasets, it is not necessary to present all the data to a learning algorithm in order to achieve a high level of accuracy. He also states that for any given dataset, there is a critical training set size, beyond which higher generalisation accuracy is not possible. Dietterich [6] has argued that overfitting does occur for large datasets, and so, training on all the data in a large dataset should be avoided.

For non-inductive learning, such as association rule mining there is also evidence to support the claims that a small subset of database records can produce results with an acceptable level of accuracy. This has been demonstrated by Zaki et al [27] and Toivonen [22].

Given the above observations, it is highly desirable to reduce the amount of data presented to data mining algorithms. This reduction should not be done on an ad hoc basis, but should be made to ensure that the probability distribution and information con-

Email: P E N Lutu patricia.lutu@up.ac.za, A P Engelbrecht engel@cs.up.ac.za

tent of the sample is the same as (or close to) that of the original dataset. Sampling is one way to reduce the amount of data presented to an algorithm. John and Langley [11] report two main approaches to sample selection that are in common use. These are static sampling and dynamic sampling. With static sampling, a random sample is drawn from the large dataset, and hypothesis testing is used to establish whether it is sufficiently similar to the parent dataset. With dynamic sampling, a decision is made after each sampled element whether to continue sampling or not. If the current sample is considered sufficient for the required level of accuracy, the sampling stops. Dynamic sampling is especially suited to incremental algorithms, such as artificial neural networks, but it is not well suited to non-incremental algorithms, such as decision trees. Static sampling, on the other hand, is applicable to any type of algorithm.

The research reported in this paper answers three questions. The first question is: *Do available statistical tests provide good guidelines for empirical estimation of sample sizes?* The second question is: *If the tests provide good guidelines, how precisely should they be used for sample size estimation?* The third question is: *How do these methods perform, compared to the theoretical guidelines for sample size estimation?* John and Langley [11] have answered the first question in part, by studying the usage of the mean test and the chi-square goodness-of-fit test. The research reported in this paper extends the investigation to the mean and variance tests and the trimmed mean test, as well as the use of information theoretic measures. To this end, experiments have been conducted to study the usage of five statistical tests in estimating sufficient sample sizes for several datasets. In order to answer the third question, a comparison has been made between the classification performance of samples evaluated empirically, and the theoretical estimates of sample complexity, based on the probably approximately correct (PAC) model of inductive learning [23]. The C5.0 classifier, which is the commercial version of the C4.5 classifier [20] has been used to establish classifier accuracy for the samples.

The main findings of this research are that first of all, testing a single sample for statistical validity, based on the mean, variance or chi-square goodness-of-fit test does not provide sufficient information about the sufficiency of the sample for classification purposes. This confirms John and Langley's [11] findings. The second finding is that testing many samples using these tests, does provide useful information about what sample sizes are very likely to be sufficient for classification in particular and data mining in general. The third finding is that in order to obtain conclusive evidence of sample sufficiency for classification purposes, information theoretic measures need to be made on the samples. The recommendation arising from this research is that hypothesis testing using the trimmed mean should be used in conjunction with an information theoretic measure in order to establish the sufficiency of samples for classification purposes. It is also shown that the recommended method results in estimates that are of more practical use, than those estimates based on the theoretical methods of PAC learning.

The rest of the paper is organised as follows. In section 2 previous work on empirical estimation of sample size as well as the theoretical estimation using PAC are discussed. Various issues related to the empirical evaluation of dataset samples are discussed in section 3. The experimental results for empirical sample evaluation and the classification results obtained from samples are presented in sections 4 and 5. Section 6 concludes the paper.

2 PREVIOUS WORK ON STATIC SAMPLING FOR INDUCTIVE ALGORITHMS

Previous work on sample size estimation for inductive algorithms is discussed in this section. The scope of the discussion is limited to those methods that statically estimate the sufficient sample size. Even though dynamic sampling has been reported to provide better estimates than static sampling, it is important to investigate static sampling methods, since dynamic sampling is not always the optimal choice for some classification algorithms. For empirical estimation, univariate hypothesis testing with the mean and chi-square goodness-of-fit tests have been reported in the literature. For theoretical estimation, the most common approach is to use PAC estimates. More recently, the Rademacher penalty [7] has been reported as a more viable alternative to the usage of PAC and the VC dimension [24]. The empirical method of univariate hypothesis testing and the PAC methods are briefly reviewed in this section. Elomaa's method [7], which is based on the Radmacher penalty is not discussed here, since it is based on dynamic sampling.

2.1 Empirical estimation of sample complexity

A random sample obtained from a large dataset, using static random sampling, should be evaluated to establish whether it is sufficiently representative of the dataset. The methods reported in the literature revolve around hypothesis testing to establish that the sample and the large dataset have the same probability distribution. John and Langley [11] discuss univariate hypothesis testing on the mean and chi-square goodness-of-fit test. For each continuous-valued attribute, hypothesis testing is done to establish whether the sample and large dataset have the same mean. For each categorical attribute, the chi-square goodness-of-fit test is used to establish whether the sample and large dataset have the same distribution.

Hypothesis testing on the mean is based on the assumption that the attribute values have a normal distribution. It is however generally known that real-life data is not always normally distributed. Skewed distributions and the presence of outliers is the norm rather than the exception. Wilcox [25] has advised that, when normality is assumed, deviations from nor-

mality result in misleading conclusions for the hypothesis testing, even for large datasets. It has also been observed in the experiments conducted for this research, that when a single random sample is drawn and found to be statistically valid, this can lead to the erroneous conclusion that all random samples of that size are statistically valid. This problem should be avoided by establishing the probability of drawing a statistically valid sample of size, say, S. If this probability is high, for example 0.95, then a claim can be made that random samples of size S are generally valid samples. Such a claim does not imply sample sufficiency. However, as the experimental results in this paper will show, sample sizes that are generally valid can be efficiently used as a starting point for identifying sufficient samples.

2.2 Theoretical estimation based on PAC and the VC dimension

The probably approximately correct (PAC) model of learning proposed by Valiant [23] and discussed by Mitchell [14], considers algorithms that learn target concepts from some concept class C, using training examples drawn at random according to some unknown, but fixed, probability distribution. PAC is concerned with the identification of classes of hypotheses that can and cannot be learned from a polynomial number of examples. PAC also defines measures of complexity of hypothesis spaces that makes it possible to define bounds for the number of training examples required for inductive learning.

PAC requires that the learner probably (with a probability of at least $1-\delta$) learn a hypothesis that is approximately correct (with predictive error ε), given computational effort and training examples that grow only polynomially with $1/\varepsilon$, $1/varepsilon$, the number of the instances,i m , and the size of the hypothesis space $|H|$. For the *agnostic (robust) learning model* (within PAC) the learner outputs the hypothesis from H that has the least error over the training data. Under this model, the learner is assured with probability $(1-\delta)$ to output a hypothesis within error ε of the hypothesis h in H, after observing m randomly drawn training examples, provided [14]

$$m \geq \frac{1}{2_{\varepsilon}^{2}}\left(\ln\frac{1}{\delta}+\ln|H|\right) \tag{1}$$

Equation (1) is applicable to hypotheses for which the size of the hypothesis space, $|H|$, is finite. For infinite hypothesis spaces, a useful measure of the complexity of H is its Vapnik-Chervonenkis dimension, $VC(H)$ [24]. $VC(H)$ is the size of the largest subset of instances that can be shattered (split in all possible ways) by H. An alternative upperbound for m under the PAC model may be restated as [14]:

$$m \geq \frac{1}{\varepsilon}\left(4\log_2\left(\frac{2}{\delta}\right)+8VC(H)\log_2\left(\frac{13}{\varepsilon}\right)\right) \tag{2}$$

One problem with using the VC dimension is that it is not always easy to estimate the VC dimension for a given classification algorithm. Additionally, the VC dimension might be infinite, as is the case for a fully grown decision tree. Auer et al [1], have derived an expression for the VC dimension of a decision tree of bounded depth. Guestrin [9] has given a bound on the size of the hypothesis space for decision trees of depth k. He has shown that for a classification problem with d attributes:

$$\log_2 H_k = \left(2^k-1\right)\left(1+\log_2 d\right)+1 \tag{3}$$

where H_k is the number of decision trees of depth k. Based on equations (1) and (3), the sample complexity for a decision tree learner of depth k, for an instance space with d attributes is:

$$m \geq \frac{\ln 2}{2_{\varepsilon^2}}\left((2^k-1)(1+\log_2 d)+1+\ln\frac{1}{\delta}\right) \tag{4}$$

The PAC model provides a worst-case estimate as it requires that the number of training examples needed, should be bounded by a single fixed-size polynomial for all target concepts and all distributions in the instance space. Haussler [10] has observed that one criticism that is often leveled at the PAC model is that the worst-case emphasis results in the estimation of the worst-case number of examples needed, and therefore makes the model unusable in practice. This is demonstrated in sections 3 and 5. Another criticism of the model is that the assumptions of well defined target concepts and noise-free data are unrealistic in practice.

3 EMPIRICAL EVALUATION OF DATASET SAMPLES

In this section, the datasets, the sampling method used and the statistical tests applied to the samples, are discussed. A distinction is made between statistical validity and sample sufficiency. This distinction is important, since it was found in the experimental results that statistical validity does not imply sample sufficiency.

3.1 The datasets used for the experiments

For the experiments, the datasets, iris, pima-diabetes, abalone, mushroom, housing16H, and adult income from the UCI machine learning repository [3] were used. For abalone, the three-class version of the dataset was used. For housing16H the attribute price was discretised into 11 intervals for classification purposes. The datasets were chosen in an effort to vary the complexity of the learning problems. The number of attributes was varied from 4 for iris to 16 for housing16H. The size of the dataset was varied as 150 for iris, 768 for pima-diabetes, 4177 for abalone, 8142 for mushroom, 22784 for housing16h and 48842 for adult income. The number of classes was varied as 2 for pima-diabetes, mushroom and adult income, 3 for abalone, 4 for iris and 11 for housing16H. Although

some of these datasets cannot be classified as large, they are sufficient in showing that smaller samples of the sets can be sufficiently representative of the whole set, for classification purposes [8]. A summary of the characteristics of the datasets is given in Table 1.

3.2 Sequential random sampling

In order to obtain random samples from the datasets, the method of sequential random sampling proposed by Jones [12] and later used by Olken and Rottem [16,17] were used. For this sampling method, the probability of each dataset record being included in the sample is uniform, and is achieved as follows. An independent uniform random variate (from the unit interval [0,1]) is generated for each record in the dataset to determine whether the record should be included in the sample or not. Suppose that a dataset of size N is to be sampled to obtain n records. If m records have already been chosen from among the first t records in the dataset, the $(t+1)^{st}$ record is chosen with probability $(n-m)/(N-t)$, where $(n-m)$ is the number of records that still need to be chosen for the sample, and $(N-t)$ is the number of records in the dataset, still to be processed.

3.3 Sufficient samples and statistical validity

A sample is considered to be statistically valid if each of its attributes has the same probability distribution as the corresponding attribute in the parent dataset. A sample is considered to be sufficient if the performance of classifiers that are constructed from it is close to that for classifiers constructed from the parent dataset. It is necessary to make this distinction, since a sample that is statistically valid is not necessarily sufficiently representative. It is argued in section 5.2 that both statistical validity and high information content are necessary for sample sufficiency.

3.4 Univariate hypothesis testing

Experiments have been conducted to compare the performance of four different statistical tests in the evaluation of samples using hypothesis testing. The four statistical tests used are based on the mean, the mean and variance, the trimmed mean and the chi-square goodness-of-fit test. For the parametric tests the null hypothesis is that the value of the parameter (mean, variance) for a sample lies within the confidence interval for the parameter in the parent dataset. The chi-square goodness-of-fit test has been applied to both categorical and continuous-valued attributes. Each continuous-valued attribute was first discretised into D intervals before applying the test. The null hypotheses tested are that the relative frequencies in the sample are the same as in the parent dataset. A sample passes the test if, for all the attributes, the null hypotheses cannot be rejected.

Hypothesis testing with the mean, variance and chi-square goodness-of-fit test are well documented in the statistics literature. Some good references are Montgomery et al [15] and Steyn et al [21]. Hypothesis testing on the trimmed mean as proposed by Yuen [26] and discussed in detail by Wilcox [25], is not that common, and needs some explanation. g% trimming, means that g% of the lowest values and g% of the highest values are ignored for purposes of hypothesis testing. The reasoning behind using the trimmed mean is to remove all possible outliers by trimming off the tails of the probability distribution, and then working with the middle part of the distribution, which is assumed to be normal. The effect of this is to reduce the variance and increase the power for hypothesis testing for non-normal distributions.

The datasets used for the experiments have continuous-valued attributes with different distributions ranging from extremely skewed, bimodal, and with the presence of extreme outliers for some attributes. The mean and variance tests are based on assumptions of normality, which are clearly violated when these datasets are used. The trimmed mean test is based on the assumption that a distribution is unimodal, so that trimming will result in hypothesis testing with the middle part of the distribution which is normal. The attributes which have bimodal distributions clearly violate this assumption.

3.5 Information theoretic measure

The entropy function, has been used as a measure of information content for the samples. The weighted average class entropy (ACE) for an attribute is computed as [20]:

$$ACE_r = \sum_{j-1}^{k} \frac{|S_j|}{|S|} H(S_j) \tag{5}$$

where:

$$H(S_j) = \sum_{i-1}^{m} P(c_i, S_j) \log_2 P(c_i, S_j) \tag{6}$$

$|S|$ is the sample size, $|S_j|$ is the size of the bin which has value j for attribute r, and, $P(c_i, S_j)$ is the number of training examples in S_j which have class label c_i.

For the computation of $H(S_j)$, categorical attributes have a natural partitioning of the attributes values. For continuous-valued attributes, binning is used at levels of 16, 32 and 64 bins for the experiments.

Given a sample S and the parent dataset D, for each sample, the total entropy for all attributes is computed as:

$$SampleEntropy(S) = \sum_{r=1}^{R} ACE_r \tag{7}$$

In the absence of a statistic that can be used for hypothesis testing, a criterion called *sufficiently close* is used. If $SampleEntropy(S)$ is sufficiently close to $SampleEntropy(D)$ for the parent, that is:

Dataset	Description	Number of attributes
Iris	4 continuous-valued attributes, 3 classes, 150 examples	4
Abalone	1 categorical attribute, 7 continuous-valued attributes, 3 classes, 4177 examples	8
Pima-Diabetes	8 continuous-valued attributes, 2 classes, 768 examples	8
Mushroom	22 categorical attributes, 2 classes, 8124 examples	22
Housing 16H	16 continuous-valued attributes, 11 classes, 22784 examples	16
Adult	6 continuous-valued attributes,	14
Income	8 categorical attributes, 2 classes, 48842 examples	

Table 1: Summary of characteristics for the datasets used

$$SampleEntropy(D) - SampleEntropy(S) < \delta \quad (8)$$

then the information content of the sample S is sufficiently close to that of the parent dataset, and the sample is considered to be sufficient for classification purposes.

4 COMPARISON OF THE SAMPLE EVALUATION METHODS

The performance measures and experimental results for the statistical tests are discussed in this section and in section 5. The first research question that was posed in section 1 is as follows: *Do available statistical tests provide good guidelines for the empirical estimation of sample sizes ?* For experimental purposes, this question was broken down as follows:

1. *Can parametric tests on the mean and variance provide informative results even when the assumptions of normality are violated ?*
2. *Can the parametric test on the trimmed mean provide informative results even when the assumptions of unimodality are violated ?*
3. *Does the goodness-of-fit test provide informative results for both categorical and continuous-valued attributes?*
4. *Do all the four tests above perform equally well?*
5. *Do information theoretic measures provide informative results for sufficient sample size estimation ?*
6. *If a sample is declared statistically valid, is it a sufficient sample for classification purposes ?*

This section provides answers to the first five questions. The fifth question is addressed in section 5.

4.1 Measuring the performance of the statistical tests

For the mean, variance and chi-square goodness-of-fit tests, performance was measured in terms of *informativeness*, as follows.

1. *If a test declares both small samples and large samples as being valid with a high probability, then that test is not very informative.*
2. *If a test declares both small samples and large samples as being valid with a low probability, then that test is not very informative.*
3. *If the probability of a sample being declared valid, monotonically increases with the sample size, then the test is said to be informative.*

In order to estimate the probability of samples of size S being declared valid, 200 samples were tested for each value of S. For each dataset, the sample size is varied at six levels : 5%, 10%, 20%, 40%, 60% and 80%. This is done in order to study the informativeness of the tests. When multiple tests are conducted using univariate hypothesis testing, it is normally necessary to make adjustments to the α values for the individual tests. One such adjustment is the Bonferroni correction. For the experimental results that are reported here, no corrections were performed and the α values that appear in tables 2, 3, and 4 are for each individual test.

4.2 Mean and variance tests

The experiments reported in this section are used to answer the question: *Can parametric tests on the mean and variance provide informative results even when the assumptions of normality are violated ?* It was stated in section 3.4, that all the datasets used have attributes that are not normally distributed.

For the means test, the hypotheses being tested are stated as follows:

H_0*: The mean value of the sample lies within the confidence interval of the mean for the parent dataset.*

H_1*: The mean value of the sample does not lie within the confidence interval of the mean for the parent dataset.*

For the variance test, the hypothesis being tested are stated as follows:

H_0*: The variance value of the sample lies within the confidence interval of the variance for the parent dataset.*

H_1*: The variance value of the sample does not lie within the confidence interval of the variance for the parent dataset.*

For the mean-only test, a sample passes the test, if, for all attributes the null hypothesis, H_0, for the mean is not rejected. For the mean-and-variance test, a sample passes the test for all attributes if the null hypotheses, H_0, for both the mean and the variance

are not rejected. Table 2 summarises the results of the hypothesis testing for the datasets.

The iris dataset has two attributes with bi-modal distributions. The abalone dataset has one attribute with very extreme outliers. For the pima-diabetes dataset, all the continuous-valued attributes have skewed (nearly-normal) distributions. The housing16H dataset has 8 attributes that are extremely skewed. The adult income dataset has two attributes (capital_loss and capital_gain) that are excessively skewed.

These attributes were ignored for hypothesis testing. It should be expected that, if the basic assumptions on which hypothesis testing is based are violated, then the results will be misleading. This is the case, for the iris and abalone datasets. One can see, from table 2, that the performance of both tests is low for both datasets, since very small samples (5%) as well as large samples are being declared valid with very high probability.

For the abalone dataset, the experiment was repeated with the two most extreme outliers removed from the height attribute. The figures in parentheses show the results (probability) of sample validity. One can conclude that the performance for the mean and variance test is improved by the removal of outliers. The informativeness of the tests is still low, even when outliers are removed.

The results for the tests on the pima-diabetes dataset indicate that, while the means-only test is very misleading, the mean-and-variance test is far more informative, than for the other two datasets. The results for the housing16H and adult income datasets indicate that the presence of extremely skewed distributions will make the mean-and-variance test totally meaningless.

The answer to the question that is posed in this section is that, while the mean-only test can be very misleading, the mean-and-variance test does provide useful information, provided that the distribution of all attributes is unimodal and there are no extreme outliers in the data, as is the case for pima-diabetes.

4.3 Trimmed mean tests

The experiments reported in this section are used to answer the question : *Can parametric tests on the trimmed mean provide informative results even when the assumptions of unimodality are violated ?* For the trimmed means tests, the hypotheses being tested are stated as follows:

H_0*: The trimmed mean value of the sample lies within the confidence interval of the trimmed mean for the parent dataset.*

H_1*: The trimmed mean value of the sample does not lie within the confidence interval of the trimmed mean for the parent dataset.*

The trimmed mean, as explained in section 3.4, ignores outliers, so that the estimate of the confidence interval of the mean is more reliable. Three levels of trimming have been used for the experiments. 5% trimming is considered to be the most conservative level. 10% trimming is a medium level, while 20% trimming is considered to be high. Wilcox [25] recommends that 20% trimming should be used to provide robustness in the presence of outliers and other deviations from normality.

For the datasets, iris, abalone, pima-diabetes and housing16H, all levels of trimming seem to provide good performance, in the presence or absence of outliers and in the presence of attributes with bi-modal distributions. Very small samples (5%) are declared valid with very low probability and the probability of a sample being declared valid, monotonically increases with the sample size. One can also see that the performance of the trimmed means tests is better than that for the mean and variance tests, as the trimmed means tests perform well for all the datasets. However, for the adult income dataset, the estimated 95% confidence intervals for the trimmed means of five of the continuous-valued attributes is so narrow that it causes problems for the hypothesis testing. In the experiments, the 95% confidence interval of the mean is estimated as: $\overline{X} \pm 1.96\sigma/\sqrt{n}$, where n is the size of the parent dataset. As the size of the parent dataset gets larger, the confidence interval estimate becomes narrower. Alternative methods of estimating the confidence interval need to be devised in order to avoid this problem.

4.4 Chi-square test for goodness-of-fit

The experiments reported in this section are used to answer the question: *Can the goodness-of-fit test provide informative results for both categorical and continuous-valued attributes?* It was stated in section 3.4, that all the datasets used have attributes that are not normally distributed. Since the chi-sqaure goodness-of-fit test is a non-parametric test, it does not make any assumptions about the distribution of the data. For this test, the hypotheses being tested are stated as follows:

H_0*: The frequency distribution of the attribute in the sample is the same as that in the parent dataset.*

H_1*: The frequency distribution of the attribute in the sample is not the same as that in the parent dataset.*

The test statistic used is:

$$\chi_0^2 = \sum_{i-1}^{k} \frac{(O_i - E_i)^2}{E^2} \tag{9}$$

where O_i is the observed relative frequency in the sample for a given attribute value, E_i is the expected relative frequency as established in the parent dataset, for the same attribute value, and k is the number of distinct values for the attribute.

This test is applied to categorical as well as continuous-valued attributes by binning (discretisation) and then establishing the frequencies for each bin. Discretisation of the continuous-valued attributes into D intervals was done at three levels of $D = 16$, $D = 32$, and $D = 64$. The reason is to establish

Dataset	**Sample size as percentage of dataset**	**Actual sample size**	**Percentage that passed the test at the significance level of 0.05**			
			Mean only		**Mean & variance**	
Iris	5%	8	91		88	
	10%	15	90		83	
	20%	30	94		87	
	40%	60	97		95	
	60%	90	98		99	
	80%	120	100		100	
Abalone	5%	209	90	(91)	44	(76)
	10%	418	91	(91)	26	(77)
	20%	835	91	(93)	12	(74)
	40%	1671	99	(96)	2	(92)
	60%	2506	100	(100)	4	(97)
	80%	3342	100	(100)	75	(100)
Pima-Diabetes	5%	38	77		17	
	10%	77	80		18	
	20%	154	86		22	
	40%	308	94		40	
	60%	462	98		65	
	80%	616	100		95	
Housing16H	5%	1139	52		0	
	10%	2278	67		0	
	20%	4556	68		0	
	40%	9112	87		0	
	60%	13668	94		0	
	80%	18224	100		0	
AdultIncome	5%	2442	76		47	
	10%	4884	88		52	
	20%	9768	96		52	
	40%	19535	96		68	
	60%	29305	100		72	
	80%	39072	100		75	

Table 2: Hypothesis testing results for the mean and variance

Dataset	**Sample size as percentage of dataset**	**Actual sample size**	**Percentage that passed the test at the significance level of 0.05**		
			5% trimming	**10% trimming**	**20% trimming**
Iris	10%	15	7	4	0
	20%	30	12	7	2
	40%	60	39	28	15
	60%	90	84	75	58
	80%	120	100	99	97
Abalone	5%	209	4	3	1
	10%	418	14	16	7
	20%	835	50	32	14
	40%	1671	70	61	50
	60%	2506	93	93	75
	80%	3342	100	100	98
Pima-Diabetes	5%	38	0	0	0
	10%	77	0	1	0
	20%	154	3	1	0
	40%	308	29	19	3
	60%	462	78	67	30
	80%	616	100	100	89
Housing16H	5%	1139	18	17	8
	10%	2278	39	25	18
	20%	4556	58	38	34
	40%	9112	83	84	63
	60%	13668	97	96	81
	80%	18224	100	100	99
AdultIncome	5%	2442	12	very narrow confidence intervals make it difficult to obtain meaningful results	
	10%	4884	36		
	20%	9768	92		
	40%	19536	100		
	60%	29305	100		
	80%	39072	100		

Table 3: Hypothesis testing results for the trimmed means tests

whether the level of discretisation affects the outcome of the hypothesis testing.

Table 4 gives the results of the hypothesis testing experiments. The level of discretisation seems to have an insignificant effect on the outcome of the hypothesis tests. This test provides good performance for the iris dataset (150 examples). Very small samples (5%) are declared valid with a very low probability, and the probability of validity increases monotonically with the sample size.

For the pima-diabetes dataset (768 examples) the results for sample sizes less than 20% are informative. However, for samples sizes above 20%, the test is less informative. For abalone dataset (4177 examples) the test has a very low performance level as both very small samples (5%) and large samples (80%) are declared valid with equally high probability. For mushroom, housing16H and adult income, the test does not provide very useful information for one to be able to distinguish between small and large samples.

It appears, from the experimental results, that this test is not powerful enough to provide useful information about samples that are larger than 200 in size. In answering the question that is posed in this section, for both categorical and continuous-valued attributes, this test does not provide informative results except for very small samples that are less than 200 in size.

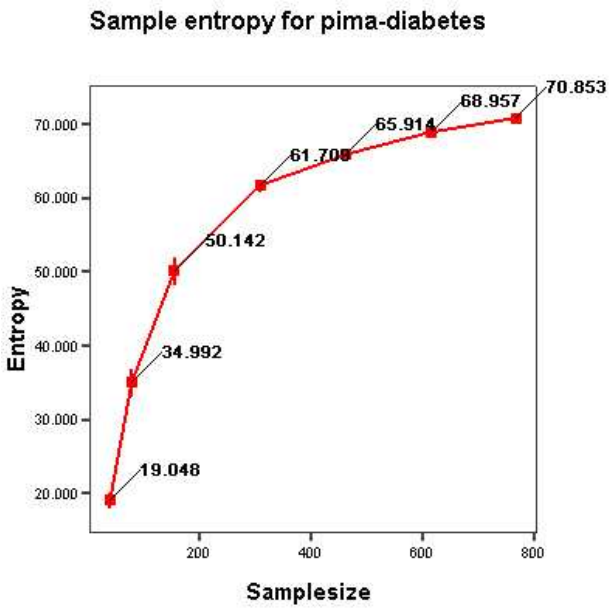

Figure 1: Sample entropy for pima-diabetes for discretisation at 64 intervals. Error bars show the 95.0% confidence interval of the mean. Dots/lines show means.

4.5 Comparison of the four tests

The fourth question that was posed is: *Do all four tests perform equally well?* From the discussion of the experimental results it should be concluded that the tests do not perform equally well. The mean-only test is not informative, and should be avoided. The mean-and-variance test will perform well if outliers are removed and all the attributes have unimodal distributions. The chi-square test for goodness-of-fit will only provide meaningful results if samples are very small: less than 200 in size. The trimmed mean test performs well in most situations investigated. It is therefore found to be the most informative test. It should however be noted that the test does not perform well in the presence of extremely skewed distributions. It should also be noted that the usage of the trimmed mean test requires attribute values to be sorted for the identification of the tails of the distribution for each attribute. This is a computational overhead which the other tests do not require.

4.6 Information theoretic measures

The fifth sub-question that was posed in section 4.1 is as follows. *Do information theoretic measures provide informative results for sufficient sample size estimation?* The information measure of equation 7 was used for measuring the information content of samples from the datasets. The abalone, mushroom and pima-diabetes datasets are used for illustration. Tables 5, 6 and 7 show the details of the measures. The information content is measured relative to the predicted attribute. Thirty measures were taken for each sample size. Continuous-valued attributes were discretised at three levels.

The symbols s and D represent the sample entropy for the sample and parent datasets respectively. $Delta\%$ is the percentage of the difference between the two values, and is a measure of how close the two values are. As can be seen from the tables the level of discretisation affects the value of the sample entropy. The higher the discretisation level, the higher the sample entropy values become.

The plots of sample entropy are shown in figures 1 and 2. As can be observed from these plots, as the sample size is increased, the information content rises steeply and then gradually flattens. This behaviour was observed for all levels of discretisation, and for the other datasets. This behaviour of the information measure can be used to provide guidelines on how sufficient sample sizes can be selected. When the plot of information content flattens, this means that increasing the sample size does not result in any significant increase in information for classification.

The sub-question that was posed in this section is whether information theoretic measures can provide useful guidelines for the estimation of sufficient sample sizes. The answer to this question is yes. Suppose we choose the discretisation level to be 64 intervals. If we set the selection criteria at say: $SampleEntropy(D) - SampleEntropy(s) <= 5\%$, then we can choose sample sizes for which the values in the last column of tables 5,6 and 7 are closest to 5%.

Dataset	**Sample size as percentage of dataset**	**Actual sample size**	**Percentage that passed the test at the significance level of 0.05 D discretisation intervals used**		
			D = 16 intervals	**D = 32 intervals**	**D = 64 intervals**
Iris	5%	8	0	0	0
	10%	15	0	0	0
	20%	30	0	0	1
	40%	60	85	94	90
	60%	90	99	100	99
	80%	120	100	100	100
Abalone	5%	209	100	99	99
	10%	418	100	100	100
	20%	835	100	100	100
	40%	1671	100	100	100
	60%	2506	100	100	100
	80%	3342	100	100	100
Pima-Diabetes	5%	38	0	0	0
	10%	77	38	58	79
	20%	154	99	98	99
	40%	308	100	100	100
	60%	462	100	100	100
	80%	616	100	100	100
Mushroom	5%	406	Discretisation levels not applicable. All values are categorical		100
	80%	6499			100
Housing16H	5%	1139	100	100	100
	80%	18224	100	100	100
AdultIncome	5%	2442	100	100	100
	80%	39072	100	100	100

Table 4: Hypothesis testing results for the chi-square goodness-of-fit tests

Sample size	**Discretisation level**					
	16 intervals		**32 intervals**		**64 intervals**	
	Mean entropy	**Delta% =100* (D-s)/D**	**Mean entropy**	**Delta% =100* (D-s)/D**	**Mean entropy**	**Delta% =100* (D-s)/D**
38	s = 19.3	35.2	20.5	54.6	19.1	73.1
77	s = 21.9	26.5	29.5	34.7	35	50.6
154	s = 24.9	16.4	35.5	21.5	50.1	29.3
308	s = 27.9	6.4	40	11.5	61.7	13
462	s = 28.3	5.0	42.8	5.3	66	7.1
616	s = 29.1	2.3	44.4	1.8	69	2.7
768	D = 29.8		45.2		70.9	

Table 5: Sample entropy measurements for pima-diabetes

Sample size	**Discretisation level**					
	16 intervals		**32 intervals**		**64 intervals**	
	Mean entropy	**Delta% =100* (D-s)/D**	**Mean entropy**	**Delta% =100* (D-s)/D**	**Mean entropy**	**Delta% =100* (D-s)/D**
209	s = 35.6	3.8	65.631	14.2	108.266	25
418	s = 35.9	3.0	71.127	6.6	126.947	12
835	s = 36.6	1.1	74.062	2.6	136.343	5.6
1671	s = 36.7	0.8	75.492	0.8	141.054	2.3
2506	s = 36.9	0.3	75.598	0.7	142.994	1
3342	s = 37.1	0	75.844	0.4	143.602	0.6
4177	D = 37.0		76.134		144.387	

Table 6: Sample entropy measurements for Abalone3C

Sample size	Mean Entropy	Delta% =100* (D-s)/D
421	s = 23.5	4.9
842	s = 24.1	2.4
1684	s = 24.5	0.8
3368	s = 24.6	0.4
5052	s = 24.7	0
6736	s = 24.7	0
8416	D = 24.7	

Table 7: Sample entropy measurements for mushroom

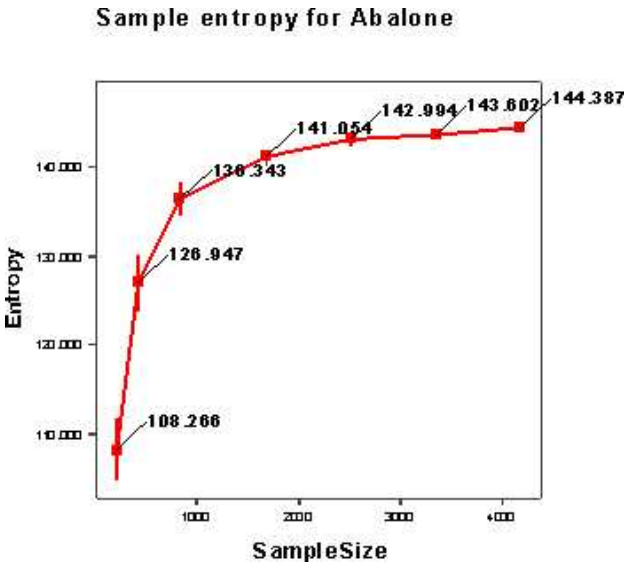

Figure 2: Sample entropy for abalone3 for discretisation at 64 intervals. Error bars show the 95.0% confidence interval of the mean. Dots/lines show means.

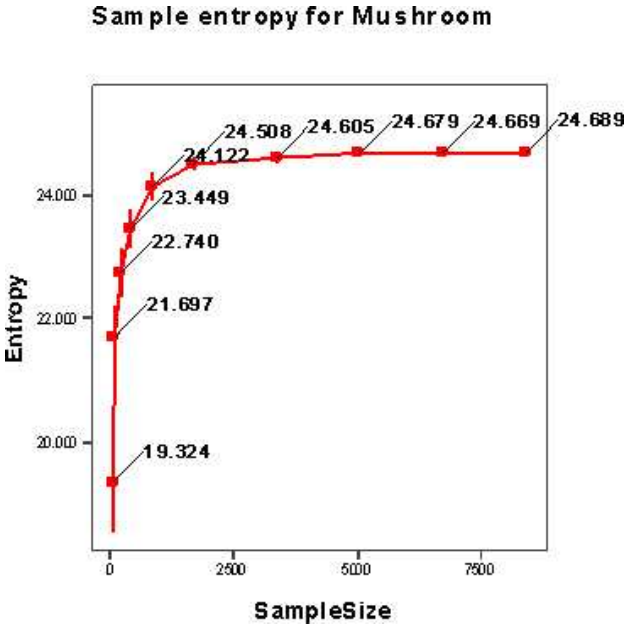

Figure 3: Sample entropy for the mushroom dataset. Error bars show the 95.0% confidence interval of the mean. Dots/lines show means.

5 RESULTS FOR DECISION TREE CLASSIFICATION

The first question that was posed for this research is as follows. *Do available statistical tests provide good guidelines for the empirical estimation of sample sizes?* This question was broken down into six sub-questions, five of which were answered in section 4. The sixth sub-question is: *If a sample is declared statistically valid, is it a sufficient sample for classification purposes ?* This sub-question is answered in this section.

The second question is: *How precisely should these methods be used for sample size estimation?* In this section, guidelines are given on how these methods could be used to estimate sufficient sample sizes.

The third question that was posed is: *How do empirical methods perform compared to the theoretical guidelines for sample size estimation ?* In the context of this question, performance is measured as follows. For a given level of accuracy, suppose that method A estimates that the sufficient sample size is S_A, and method B estimates that the sufficient sample size is S_B. If $S_A < S_B$ then the conclusion is that method A has a higher performance than method B. In this section the theoretical estimates using PAC are compared with empirical estimates obtained with hypothesis testing using the four statistical tests.

5.1 Measuring classification performance

The C5.0 decision tree algorithm was used to construct classifiers that were used to evaluate the predictive performance for different sample sizes. This was done to establish whether sample validity implies sample sufficiency. For each sample size, S from 5% to 80% of the dataset, thirty (30) samples were selected at random, classifiers were constructed and the predictive accuracy of each classifier was measured using 10-fold cross-validation. The mean accuracy for the thirty classifiers, and the confidence interval of the mean were then plotted for each sample size.

In the context of comparing two samples for classification performance, the sample which provides a classifier with the higher predictive accuracy, is considered to provide the better perormance. Furthermore, a sample which provides a level of predictive accuracy that is not statistically significantly different from that provided by the parent dataset, is considered to be a sufficient sample.

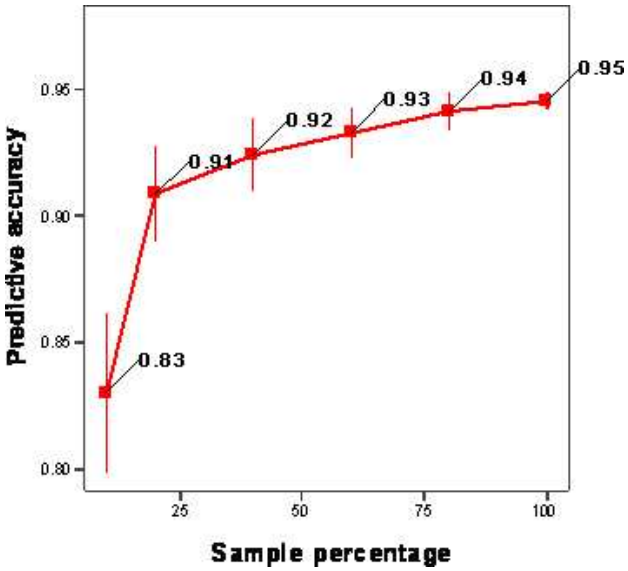

Figure 4: Predictive accuracy for the Iris dataset. Error bars show the 95.0% confidence interval of the mean. Dots/lines show means.

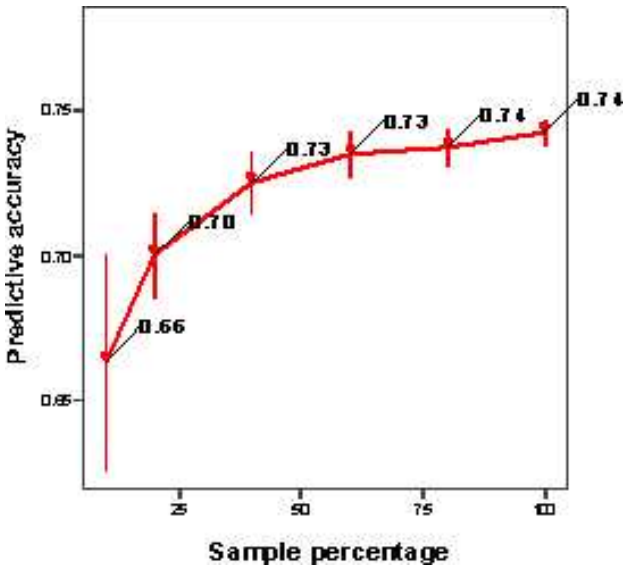

Figure 5: Predictive accuracy for the Pima-diabetes dataset. Error bars show the 95.0% confidence interval of the mean. Dots/lines show means.

5.2 Results for C5.0 classification

Six sets of classifiers were constructed: one set for each of the datasets. Figures 4, 5, 6, 7 and 8 show the plots of predictive accuracy for the classifiers. Figure 4 illustrates that the predictive accuracy for the iris samples begins to flatten from sample sizes of 20%. Figure 5 illustrates that for the pima-diabetes dataset, the flattening begins at 40%. Figure 6 illustrates that for the three-class abalone dataset, the predictive accuracy begins to flatten at sample sizes of 80%. For the housing16H dataset, the curve begins to flatten for sample sizes of 60%. For adult income the curve begins to flatten for sample sizes of 40%. The accuracy of mushroom was found to be 93% for 0.5% samples (42) examples, 99% for 1% samples (84 examples) and 100% for 5% samples (421 examples) and was therefore not plotted. When a learning curve is flat (or nearly flat) then increasing the sample size does not significantly affect the classification accuracy.

It is useful to establish whether the results of the experiments on hypothesis testing for sample validity could provide practical guidelines as to what sample sizes provide classifier accuracy in the region where the learning curve begins to flatten, or is actually flat. Looking at the results of section 4, for the mean-and-variance test and the trimmed mean test, this happens for sample sizes where almost any random sample (95% chance) that is drawn from the dataset, is found to be statistically valid.

In section 3.3, a distinction was made between sample sufficiency and statistical validity. A sample size is considered to be sufficient, if the classifiers constructed using that sample size have a predictive accuracy that is not statistically significantly different from those constructed with the parent dataset. In order to establish whether a sample size is sufficient, hypothesis testing was done to establish, whether this is the case, for different sample sizes. The sample sizes

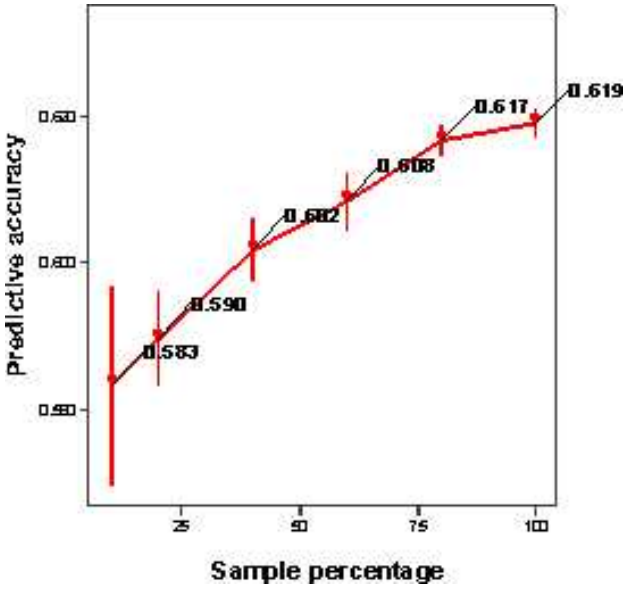

Figure 6: Predictive accuracy for the Abalone3C dataset. Error bars show the 95.0% confidence interval of the mean. Dots/lines show means.

chosen were those for which the hypothesis testing results indicated at least a 95% chance of obtaining a valid sample. The rational behind this decision is that, probabilistically, a small chance of not being able to obtain a valid sample, should be allowed. Additionally, this high probability is observed for samples sizes that are large relative to the whole dataset. The information content for these samples should therefore be close to that of the whole dataset.

The Student's t-test was used to test the following hypothesis for each sample size.

H_0: *The mean accuracy for the sample size is equal to the mean accuracy for the parent dataset.*

H_1: *The mean accuracy for the sample size is not*

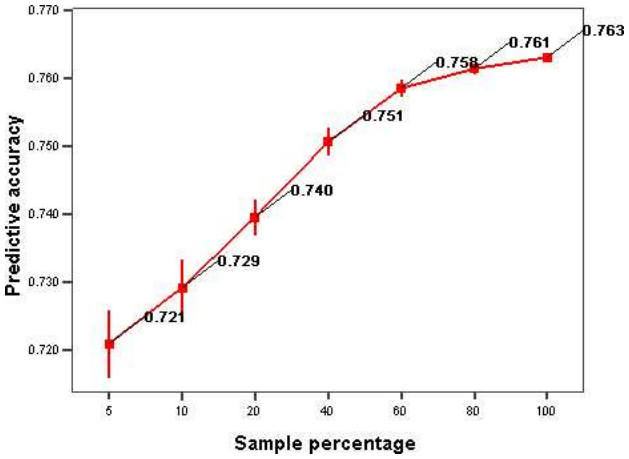

Figure 7: Predictive accuracy for the Housing16H dataset. Error bars show the 95.0% confidence interval of the mean. Dots/lines show means.

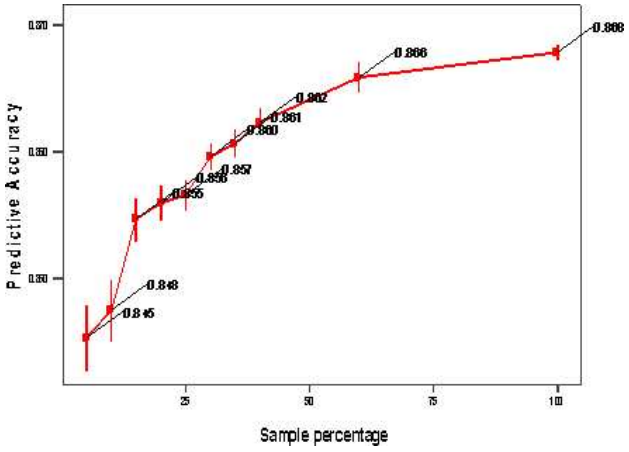

Figure 8: Predictive accuracy for adult income dataset. Error bars show the 95.0% confidence interval of the mean. Dots/lines show means.

equal to the mean accuracy for the parent dataset.

Tables 7, 8, 9, 10, 11 and 12 show the results of the hypothesis testing. The first column shows the sample size as a percentage of the dataset size, and the actual sample size in parenthesis. The abbreviations in the second column should be interpreted as: M-O (mean only test), M-V (mean and variance test), M-T (trimmed means tests) and G-O-F (chi-square goodness-of-fit-test). The last column in each table indicates whether the difference between the means is considered to be zero (equal means) or not. When the confidence interval (CI) of the difference of the means includes zero, the means are considered to be equal.

One of the sub-questions posed in section 4.1 is : *If a sample is declared statistically valid, is it a sufficient sample for classification purposes ?* The answer to this question is, no, for the following reasons . For all the datasets, sample sizes of 40% or less provide accuracy which is significantly less than for the whole dataset, at the significance level of $\alpha = 0.05$. This leads to the conclusion that statistical validity does not imply sample sufficiency. The results of tables 2, 3 and 4 of section 4 illustrate that for all four tests, small samples of up to 40% in size are being declared statistically valid. The evidence of tables 8,9,10,11 and 12 illustrates that for the datasets used, samples of size 40% or less are not sufficient for classification purposes.

The results of table 8 illustrate that for the Iris dataset samples sizes of 80% (120 examples) provide accuracy which is the same as that for the whole dataset. For the abalone dataset, the results of table 9 illustrate that sample sizes of 80% (3342 examples) provide the same accuracy as the whole dataset. For the Pima-diabetes dataset, the results of table 10 illustrate that 60% of samples (462) provide the same accuracy as the whole dataset. For the housing16H dataset, the results of table 11 illustrate that 60% of samples (13668) provide the same accuracy as the whole dataset.These results indicate that statistical validity should be augmented with others measures in order to obtain useful results for sample estimation. The measure that one can identify from these experiments is information content.

5.3 Guideline for determining a sufficient sample size

In section 4.5, it was argued that the trimmed means test and the mean and variance test (in the absence of outliers) are useful in establishing a sample size where any random sample that is drawn from a parent dataset is generally statistically valid. In section 4.6, it was demonstrated that information theoretic measures do provide a method of establishing sufficient sample sizes. Table 13 is used to summarize the results.

When selecting a sufficient sample size, it is recommended that both types of tests should be used. The less computationally intensive hypothesis tests should be used first to establish a sample size where the probability of selecting a valid sample is high. For example, a probability of 0.7 (70%) could be used. The sample entropy (or other information measures) should then be used, starting with samples of this size to establish a sample size for which the information content is close to that of the parent dataset.

5.4 Comparison of theoretical and empirical estimates

The second question that was posed in section 1 is: *How do the empirical methods discussed perform, compared to the theoretical guidelines for sample size estimation?* A comparison is made between the PAC theoretical estimates for sample complexity, and the empirical estimates. This is done only for the Pima-diabetes dataset since it is a concept learning problem (only two classes), for the concept '*diabetes*'. Table 14 gives the sample complexity as estimated by the proposed empirical method suggested in section 5.3 and the PAC theoretical estimates. For the PAC estimates, equation 4 is used. This is done to give some idea of the type of estimates that PAC provides. The

Sample size Percent (actual)	**Declared sufficient by test**	**Mean accuracy for sample (standard deviation**	**Mean accuracy for dataset (standard deviation**	**Student's T test for equality of means (equal vars not assumed)**	
				95% CI of difference of means	**Means are equal**
40% (60)	M-O	0.924 (0.039)	0.945 (0.009)	Lower: -0.0358 Upper: -0.0062	no
60% (90)	M-O M-V G-O-F	0.933 (0.027)	0.945 (0.009)	Lower: -0.0229 Upper: -0.0017	no
80% (120)	M-O, M-V, T-M, G-O-F	0.941 (0.020)	0.945 (0.009)	Lower: -0.0121 Upper: 0.0041	yes

Table 8: C5.0 classifier accuracy for the Iris dataset

Sample size Percent (actual)	**Declared sufficient by test**	**Mean accuracy for sample (standard deviation**	**Mean accuracy for dataset (standard deviation**	**Student's T test for equality of means (equal vars not assumed)**	
				95% CI of difference of means	**Means are equal**
40% (1671)	M-O G-O-F	0.602 (0.118)	0.619 (0.005)	Lower: -0.0219 Upper: -0.0123	no
60% (2506)	M-O M-V G-O-F G-O-F	0.608 (0.108)	0.619 (0.005)	Lower: -0.0149 Upper: -0.0060	no
80% (3342)	M-O, M-V, T-M, G-O-F	0.617 (0.006)	0.619 (0.005)	Lower: -0.005 Upper: 0.007	yes

Table 9: C5.0 classifier accuracy for the Abalone3C dataset

Sample size Percent (actual)	**Declared sufficient by test**	**Mean accuracy for sample (standard deviation**	**Mean accuracy for dataset (standard deviation**	**Student's T test for equality of means (equal vars not assumed)**	
				95% CI of difference of means	**Means are equal**
20% (154)	G-O-F	0.700 (0.004)	0.742 (0.013)	Lower: -0.0577 Upper: -0.0268	no
40% (308)	M-O G-O-F	0.725 (0.030)	0.742 (0.013)	Lower: -0.0293 Upper: -0.0049	no
60% (462)	M-O M-O G-O-F G-O-F	0.735 (0.0228)	0.742 (0.013)	Lower: -0.0172 Upper: -0.0021	yes
80% (616)	M-O, M-V, T-M, G-O-F	0.737 (0.018)	0.742 (0.013)	Lower: -0.0134 Upper: 0.0029	yes

Table 10: C5.0 classifier accuracy for the Pima-diabetes dataset

Sample size Percent (actual)	Declared sufficient by test	Mean accuracy for sample (standard deviation	Mean accuracy for dataset (standard deviation	Student's T test for equality of means (equal vars not assumed)	
				95% CI of difference of means	Means are equal
40% (60)	GOF	0.751 (0.006)	0.763	Lower: -0.016 Upper: -0.015	no
60% (90)	GOF M-O T-M	0.758 (0.004)	0.763	Lower: -0.007 Upper: -0.006	no
80% (120)	M-O, GOF M-O T-M	0.761 (0.002)	0.763	Lower: -0.0004 Upper: 0.0008	yes

Table 11: C5.0 classifier accuracy for the housing16H dataset

Sample size Percent (actual)	Declared sufficient by test	Mean accuracy for sample (standard deviation	Mean accuracy for dataset (standard deviation	Student's T test for equality of means (equal vars not assumed)	
				95% CI of difference of means	Means are equal
40% (19536)		0.862 (0.001)	0.868 (0.001)	Lower: -0.006 Upper: -0.004	no
60% (29304)		0.866 (0.002)	0.868 (0.001)	Lower: -0.003 Upper: -0.001	yes
80% (39037)		0.867 (0.001)	0.868 (0.001)	Lower: -0.001 Upper: 0	yes

Table 12: C5.0 classifier accuracy for the adult income dataset

Dataset	Sample percentage for which Delta% ¡= 5% (for 64 levels of discretisation)	Probability of obtaining a valid sample with the T-M test for 5% trimming
Iris	60%	84%
Pima-diabetes	80%	78%
Abalone	40%	70%

Table 13: Comparison of hypothesis testing and sample entropy results

value of k, the tree depth, is set to 10, and the number of attributes d is 8. One can see that in general, the PAC estimates are much higher than those obtained with the suggested empirical method. Since the dataset for pima-diabetes is 768 in size, it is not possible to provide an empirical estimate for an accuracy of 95%, using the empirical method presented. Linear regression and extrapolation could however be used to obtain this estimate.

6 CONCLUSIONS AND SUGGESTIONS FOR FUTURE WORK

The first question that was posed is: *Do available statistical tests provide good guidelines for the empirical estimation of sample sizes ?* A number of conclusions can be drawn from the discussion of sections 4 and 5. First of all, testing a single sample for statistical validity, based on the mean, variance or chi-square goodness-of-fit test does not provide sufficient information about the sufficiency of the sample for classification purposes. The second conclusion is that the chi-square goodness-of-fit test does not provide useful information when sample sizes are large. The third conclusion is that, testing many samples using the mean-and-variance or trimmed-means test, does provide useful information about what sample sizes are very likely to be sufficient for classification. Care should be taken not to use the mean-and-variance tests in the presence of outliers and extremely skewed distributions.

The second question that was posed is: *If the tests provide good guidelines, how precisely should they be used for sample size estimation ?* The fourth conclusion that can be drawn from the experimental results is that in order to obtain conclusive evidence of sample sufficiency for classification purposes, information theoretic measures need to be made on the samples. The recommendation arising from this research is that hypothesis testing should be used in conjunction with an information theoretic measure in order to establish the sufficiency of samples for classification purposes.

The third question that was posed is: *How do these methods perform, compared to the theoretical guidelines for sample size estimation?* It can also be concluded that the recommended method results in estimates that are of more practical use, than those estimates based on the theoretical methods of PAC learning theory.

In future research the following issues will be addressed. The methods used to estimate confidence intervals for hypothesis testing become problematic as dataset sizes increase. Other methods for this purpose will be investigated. For categorical attributes, a better test needs to be used for hypothesis testing on large datasets. Continuous-valued attributes with skewed distributions can cause problems for the tests that were investigated. Appropriate methods for dealing with these types of data need to be further investigated. The datasets used in the experiments are not very large. The experiments will be conducted on much larger datasets. Finally, one information theoretic measure was tested. Investigation of other information theoretic measures is the subject of current research.

REFERENCES

[1] Auer, P., Holte, R., and Maas, W. 1995. "Theory and application of agnostic PAC learning with small decision trees". *Proc. Twelveth International Conference on Machine Learning*, pp 21-29.

[2] Bishop, C.M. 1995. "Neural Networks for Pattern Recognition", Oxford University press.

[3] Blake, C.L. & Merz, C.J. 1998. "UCI Repository of machine learning databases", University of California, Irvine, Department of Computer Science. URL `http://www.ics.uci.edu/m̃learn/MLRepository.html`

[4] Catlett, J. 1991. "Megainduction: a test flight." *Proc. Eighth Workshop on machine Learning*. Morgan Kaufmann, 1991, pp 596-599.

[5] Cover, T..M. and Hart, P.E., 1967. "Nearest Neighbor Pattern Classification". *IEEE Transactions on Information Theory*, vol. IT-13, No.1, 1967, pp.21-27

[6] Dietterich, T., 1995. "Overfitting and undercomputing in machine learning". *ACM Computing Surveys*, vol 27. no. 3, pp 326-327.

[7] Elomaa, T. and Kaariainen M., 2002. "Progressive Rademacher sampling". *Proc. 18th National Conference on Artificial Intelligence*. AAAI-2002, Edmonton, Canada. pp140-145.

[8] Engelbrecht, A.P., 1999. "Sensitivity Analysis of Multilayer Neural Networks". PhD thesis, University of Stellenbosch, South Africa.

[9] Guestrin, C. 1995. "PAC-learning, VC dimension and margin-based bounds". Lecture notes, Carnegie Mellon University. URL `http://www.cs.cmu.edu/afs/cs.cmu.edu/usr/guestrin/www/Class/10701/slides/pac-vc.pdf`

[10] Haussler, D., 1990. "Probably approximately correct learning". *Proc. Eighth National Conference on Artificial Intelligence*. AAAI/MIT Press, pp1101-1108.

[11] John, G.H, and Langley, P., 1996. "Static versus dynamic sampling for data mining". *Proc. Second International Conference on Knowledge Discovery in Databases and Data Mining*. AAAI/MIT Press, pp367-370.

[12] Jones, T., 1962. "A note on sampling from tape files". *Communications of the ACM*, vol 5, no 1, pp343.

[13] Kohavi, R., 1996. "Scaling up the accuracy on Naïve-Bayes classifiers: a decision tree hybrid". *Proc. Second International Conference on Knowledge Discovery and Data Mining*, AAAI Press, pp202-207.

[14] Mitchell T.M., 1997. *Machine Learning*. McGraw-Hill.

[15] Montgomery,D.C, Ronger G.C and Hubele N.F, 2004. *Engineering Statistics*. Wiley, New York.

[16] Olken, F. and Rotem, D., 1995. "Random sampling from databases - A survey". (invited paper), *Statistics and Computing*, March 1995, vol.5, no.1, pp 25-42, Chapman Hall.

Classifier performance		**Sample complexity**	
Accuracy	Error rate	Theoretical estimate using PAC ($\delta = 0.01$)	Empirical estimate using suggested method
0.66	0.34	12 285	77
0.70	0.30	15 779	154
0.73	0.27	19 480	308
0.74	0.26	21 008	616
0.95	0.05	568 048	Not enough data for this estimate

Table 14: Empirical and theoretical sample complexity for Pima-diabetes

[17] Olken F., 1993. "Random Sampling from Databases". PhD thesis, University of California at Berkeley.

[18] Palmer C.R. & Faloutsos, C., 2000. "Density biased sampling: An improved method for data mining and clustering", Proceedings of the ACM SIGMOD Conference, 2000, pp82-92.

[19] Provost, F., Jensen, D., and Oates, T., 1999. "Efficient progressive sampling". *Proc. Fifth ACM DIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, pp 23-32.

[20] Quinlan, J.R., 1993. *C4.5: Programs for Machine Learning*. Morgan Kauffman, San Francisco.

[21] Steyn, A.G.W., Smit, C.F., du Toit, S.H.C. and Strasheim, C. 1996. *Modern Statistics in Practice*. Van Schaik Academic, Pretoria.

[22] Toivonen, H., 1996. "Sampling large databases for association rules". *Proc. Twenty-second Conference on Very Large Databases - VLDB96*, Mumbai India, pp134-145

[23] Valiant, L.G. 1984. "A theory of the learnable". *Communications of the ACM*, vol27, no. 11, pp1134-1142.

[24] Vapnik, V.N. 2000. *The Nature of Statistical Learning Theory*, Second Edition. Springer, New York.

[25] Wilcox, R.R. 2001. *Fundamentals of Modern Statistical Methods: Substatially Improving Power and Accuracy*. Springer-Verlag, New York.

[26] Yuen, K.K. 1974. "The two sample trimmed t for unequal population variances". *Biometrica*, 61 pp165-170.

[27] Zaki, J.M., Parthasarathy, S., Li, W., and Ogihara, M., 1997. "Evaluation of Sampling for Data Mining of Association Rules". *7th International Workshop on Research Issues in Data Engineering (RIDE–in conjunction with ICDE)*, pp 42-50, Birmingham, UK.