

Accuracy Estimation with Clustered Dataset

Ricco Rakotomalala*

Jean-Hughes Chauchat*

Francois Pellegrino**

*ERIC Laboratory - University of Lyon 2
Lyon - France

Email: {Ricco.Rakotomalala, Jean-Hughes.Cauchat}@univ-lyon2.fr

**Laboratoire Dynamique du Langage - University of Lyon 2
Lyon - France

Email: Francois.Pellegrino@univ-lyon2.fr

Abstract

If the dataset available to machine learning results from cluster sampling (e.g. patients from a sample of hospital wards), the usual cross-validation error rate estimate can lead to biased and misleading results. An adapted cross-validation is described for this case. Using a simulation, the sampling distribution of the generalization error rate estimate, under cluster or simple random sampling hypothesis, are compared to the true value. The results highlight the impact of the sampling design on inference: clearly, clustering has a significant impact; the repartition between learning set and test set should result from a random partition of the clusters, and not from a random partition of the examples. With cluster sampling, standard cross-validation underestimates the generalization error rate, and is deficient for model selection. These results are illustrated with a real application of automatic identification of spoken language.

Keywords: Accuracy estimation, Supervised Learning, Clustered dataset.

1 Introduction

Most of the time, learning is organized on a dataset which is a mere sample taken from the universe to which the results are to be generalized. Concerning a supervised learning task, measuring the quality of the generalization is known as the "assessment" (Stone 1974). Several measures of the quality exist (Lavrac 1999) (generalization error rate, sensitivity, specificity, ROC curve, ...). They are usually obtained through resampling methods which are often applied under the hypothesis that the learning set is a simple random sample of observations (independent and identically distributed - *iid* - observations) from the universe of interest (Efron & Tibshirani 1995).

In practice, this hypothesis is rarely verified; the available dataset is often the result of cluster sampling, or (more generally) two-stage sampling:

- patients from a sample of hospital wards;
- X-ray images, from various angles, of a sample of patients;
- children from a sample of classrooms or schools;

- land samples from an oil drilling rig sample...

From these examples, one understands that access to individuals has been only possible through the cluster. For instance, the patients cannot be selected one by one; a sample of hospitals is selected, in which all patients (cluster sampling) or a sub-sample (two stage sampling) are observed. Clusters are not the result of some computation on the available data. The clustering structure is inherent to the data, part of data collection, "meta-information" needed to understand and to use the data.

In this paper, standard cross-validation results (under simple random sampling assumption) are shown to be overly optimistic when the dataset is actually clustered. A modification to the cross-validation procedure that accounts for the sampling design is suggested, according to the usual applications of resampling methods in sample survey problems (Shao & Tu 1995). The proposed modification is supported, first, by an application to simulation data and secondly, by an application to speech recognition.

In section 2, an adaptation of cross validation to cluster sampling will be examined. In section 3, using a simulation, the sampling distribution of the generalization error rate, under cluster or simple random sampling, will be compared to the respective true values; the consequences of model selection will be examined. In section 4, results of an application to a real life database will be presented: the problem is to automatically recognize the language spoken by a sample of individuals using the physical analysis of the audio signal of their voices. Related work is reviewed in section 5. Lastly, a conclusion and future works are presented in section 6.

2 Adapting cross-validation to cluster samples

The true error rate (Err) is a measure of how accurately the classification, built with the learning sample, would be if they were applied to the whole universe. As the universe cannot be observed completely, a learning set (the sample) is used to infer about the universe, and only an estimate of the error rate (\hat{Err}) can be computed. In this case, the sample selection is an integral part of the inference process, and any evaluation should account for it.

In reality, survey design information is seldom made available, notably absent on the benchmark datasets available on Internet servers (for example, those from the UCI repository (Bay 1999)). In fact, statistics and methods proposed to measure the error rate rely on a simple random assumption, hardly ever realized.

The usual cross validation estimate (under *iid*, independently and identically distributed data, hypoth-

Copyright ©2006, Australian Computer Society, Inc. This paper appeared at Australasian Data Mining Conference (AusDM 2006), Sydney, December 2006. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 61. Peter Christen, Paul Kennedy, Jiuyong Li, Simeon Simoff and Graham Williams, Ed. Reproduction for academic, not-for profit purposes permitted provided this text is included.

esis) of the error rate (Err) and the adaptation to cluster sampling are described in this section.

2.1 Cross-validation on *iid* samples

The cross-validation method (in J folds) usual to machine learning proceeds as follows (Stone 1974), supposing *iid* observations:

1. a sample of n individuals are obtained from the population;
2. the n cases on the learning set are randomly split in J folds of size n_j , usually $n_j = \frac{n}{J}$;
3. the learning algorithm is applied to the whole data set, save part j ;
4. the rules learned in (3.) are applied to the cases left out, fold j , and an error rate $\hat{Err}_{(j)}$ is measured;
5. the "generalization error rate" Err is estimated as

$$\hat{Err} = \sum \frac{n_j}{n} \hat{Err}_{(j)}$$

The estimator \hat{Err} is biased: $E(\hat{Err}) > Err$ because the samples used for learning in (2.) are approximately size $n \frac{(J-1)}{J} < n$; the bias decreases as J increases (of course with *iid* samples), but the random variation of \hat{Err} and computation time grow with J . But, in spite of these drawbacks, it is proved that in the particular case of leave-one out cross-validation for instance, the worst-case error of this estimate is not much worse than that of the training error estimate (Kearns & Ron 1997).

2.2 Cross-validation on cluster samples

If the learning set was obtained through cluster sampling, then the standard procedure must be modified as follows:

1. A sample of n individuals, grouped in G clusters of respective size n_g observations ($g = 1, \dots, G$);
2. The G clusters are subdivided in J parts, part j thus comprising G/J clusters, with $n_j = \sum_{g \in j} n_g$ observations;
3. Steps (3.) to (5.) of the standard procedure are then applied.

There is a "cluster effect" when the variability within the clusters is small compared to that of the whole population; then, for a given sample size n , the true generalization error rate increases. This must be apparent in the cross-validation estimation process.

3 Application on simulated data

The main interest of a simulation model is that the true error rate is known. In effect, one can either compute the theoretical distribution-based error rate, or, using a random number generator, create as many individuals as needed to construct the test data set and estimate the error rate with controlled precision. The latter option was chosen for this paper.

The example described here uses a decision tree learning algorithm and two explanatory variables (for an easier interpretation of the charts). We will deepen our analysis later (sec. 3.3) by studying the effect of increasing number of attributes and the choice of the learning algorithm.

3.1 The simulation model

For this problem, the objective of the learning algorithm is to distinguish between two classes: positives (+) and negatives (o). In the universe of reference, individuals are grouped in clusters of size $2 \times m$, of which

m are positive ("+") and m are negative ("o"). The positives, independently their cluster, are distributed according to a bivariate normal distribution with zero mean and $s^2 \times I$ covariance matrix, where s is a constant and I is an identity matrix. The negatives of a cluster are also normally distributed with the same covariance matrix but their mean is located on the circle of radius 1. The (negatives) cluster means are uniformly and randomly distributed on the unit circle (for example figure 1.a). The dataset contains g clusters, that is, $n = 2 \times m \times g$ individuals.

Hence there are three parameters: s , the dispersion of each half cluster about its mean, m , the cluster size for one class value, and g , the number of clusters on the learning set.

For each value of these parameters ($s = 0.1, 0.2, 0.5, 1$; $m = 5, 10, 20, 40$; $g = 10$), 100 learning sets were randomly generated, as well as a testing set of 1,000 clusters for the precise estimation of the true generalization error rate.

We used the C4.5 decision trees algorithm (Quinlan 1993), that can approximate any linear or non-linear boundary with broken lines made up of segments parallel to either axis.

For each learning sample (Figure 1):

1. the decision trees were constructed using the learning set;
2. the "true" error rate is approximated by the error of the tree model on the large test set (Figure 1.d);
3. the cross-validation error rate estimate was computed by taking the clustering into account (sec. 2.2);
4. the cross-validation error rate estimate was computed as if the data were *iid* (disregarding the clustering).

Let's review the algorithm using the example in figure 1.a which comprises 10 clusters, that is $n = G \times 2 \times m = 10 \times 2 \times 10 = 200$ examples. A tree (figure 1.b) is constructed on that base and the error rate is zero (figure 1.c); estimation of the true error rate, on the large dataset, is shown figure 1.d.

For our experimentation, we choose a $J = 10$ clustered cross-validation, it makes a good bias-variance trade-off of the error evaluation (Dietterich 1998). Because we have 10 clusters in the synthetic dataset, one cluster is removed at each step of our clustered cross-validation. One of those steps is depicted in figure 2: when learning on the 9 leftmost clusters, the resulting tree (figure 2.a) poorly classifies 6 of the 10 negatives set aside (figure 2.b).

3.2 Simulation results

Many interesting elements can be underlined (Figure 3):

- the true error rate Err increases with s , the relative cluster variability, and decreases with m , the cluster size;
- standard cross-validation, disregarding the cluster effect, severely underestimates the error rate;
- estimation bias increases with the cluster effect: bias is maximum when $s = 0$, that is when all individuals are identical ($s = 0.1$ and $s = 0.2$ on figure 3);
- cross-validation accounting for cluster effect slightly overestimates the true error rate; as mentioned earlier, this was expected because the cross-validation uses, at each step, a fraction of the available sample to construct the prediction model.

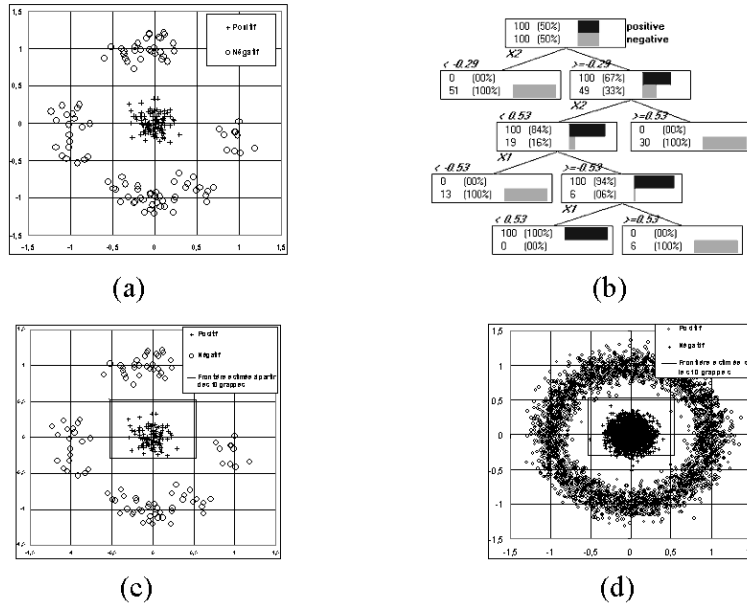


Figure 1: Learning on a sample of 10 clusters (a, b, c), applying classification rules on the generated test set (d)

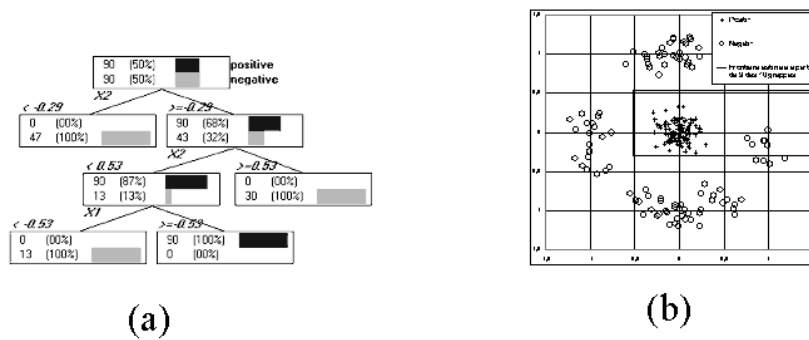


Figure 2: One step of clustered cross-validation

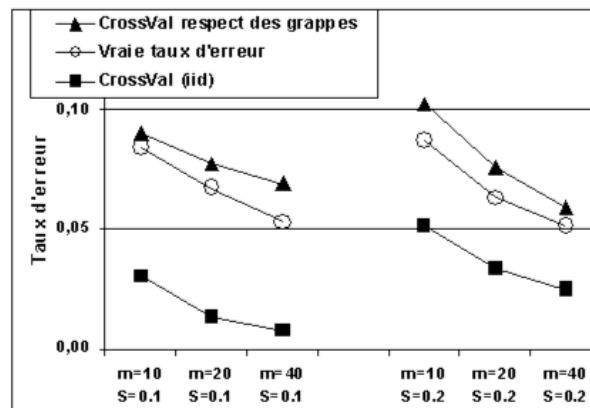


Figure 3: True and estimated error rate for $G = 10$ clusters: $s = 0.1$, $s = 0.2$ et $m = 10, 20$ et 40 . Average on 100 simulations for each case.

3.3 Simulations using alternative algorithms and multiple attributes

The preceding simulations illustrated cross-validation with clustered data, and some reasons why clustering must be accounted for. In this section, the error estimation bias is examined under varying experimental set-ups :

- increased number of predictive attributes : $dim = 2, 3, 5, 10$;
- different learning algorithms : C4.5 Decision Tree, 1-Nearest Neighbour, Naive Bayes (Hastie, Tibshirani & Friedman 2001).

The simulation population is similar to the previous one, but here the negatives (o) are centered on a random point of the hyper-sphere of dimension $dim = 2, 3, 5$ or 10 .

3.3.1 General results for each learning algorithm

Experiments confirm the results of section 3.2; whatever the learning algorithm and whatever the dimension of the space:

- standard cross-validation (assuming iid observations) always under-estimates the true error rate;
- cross-validation accounting for clusters is much closer to the true error rate.

That is :

- With the decision tree (C4.5), standard cross-validation (iid) severely under-estimates the error rate, regardless of the sample size ($n = 10 \times 2 \times m$), the clustering effect (here, noted s), and the dimension of the space;
- The nearest neighbour method (1-NN) is very sensitive to the clustering effect : if, in each class, members of a given cluster are closer to one another than those of other clusters ($s = 0.1$), then standard cross-validation dramatically under-estimates the error rate by giving an estimate of zero; notably, the true error rate with 5 predictive attributes is about 15%; and more than 30% with 10 predictive attributes. The bias of the standard method is important even with moderate clustering effect ($s = 0.3$), especially when many predictive attributes are at play ($dim = 5$ or 10). When clustering effect is smaller ($s = 0.5$) and many predictive attributes are used, the 1-NN method learns almost nothing (the true error rate increases to 50% for two classes, even with samples of size $n = G \times 2 \times m = 10 \times 2 \times 50 = 1000$ cases) yet the standard method remains "optimistic", and the bias increases with the sample size.
- With Naive Bayes algorithm, the three error estimates are almost identical when clustering effect is important ($s = 0.1$) : it is a peculiar interaction of our synthetic data and the Naive Bayes estimator. In effect, this method discretizes continuous attributes into intervals before learning. Here, because the low variance clusters of negatives orbit around a kernel of positives, the pre-processing of each attribute always give the correct three-class discretization, the positives in the center class flanked by classes of negatives. This artifact decreases when the variance increases, causing positives and negatives to intertwine and classes to be less homogeneous. Under-estimation (under the iid assumption) increases when clustering decreases (larger s), and increases in sample size and number of attributes.

Language	Speaker	Rec./Speaker	Avg.length/Rec.
German	10	20	21,9
English	10	15	17,6
Spanish	10	15	20,9
French	10	10	21,9
Italian	10	15	21,7
TOTAL	50	750	

Table 1: MULTEXT dataset, cluster structure.

3.3.2 Model selection

The results are summarized in Fig. 4 which shows that standard cross-validation can yield to very poor results if the data set is clustered, and if the clustering effect is significant ($s = 0.1$):

- In this case, judging by the standard *iid* cross-validation (figure 4.a), the nearest neighbour (1-NN) is always better than the other strategies because its error rate is always zero, though the true error rate (figure 4.c) may be larger than that of naive Bayes as soon as at least three attributes are involved ; this last result is confirmed by cross-validation under clustering (figure 4.b);
- Still under strong clustering ($s = 0.1$), the decision trees C4.5 seem better than Naive Bayes in every situation, as indicated by standard cross-validation (figure 4.a). This is deceiving, for the Naive Bayes method outperforms the three other methods (figure 4.b and 4.c) when at least three predictive attributes are present (see 3.3.1).

The simulation is, of course, particular. Still, depending on the data, standard cross-validation can lead to a very poor choice of the learning method.

4 Application to real data: speech recognition

4.1 Statement of the problem

Language identification from sound bites is an emerging domain of automatic speech processing. In this era of international and global media, the stakes are numerous, be it man-machine interface or computer-assisted human dialogue.

Most of the approaches developed so far use statistical modelling of phonetic (nature of sounds) and phonotactic (how the sounds are assembled) characteristics of the various languages (Zisman & Berkling 2001). Such approaches require vast amounts of sound recordings along with their phonetic transcriptions (entirely supervised learning).

Data mining techniques, with innovative parameterization, can give convincing results with partially supervised learning and smaller learning data sets.

4.2 The task and the data

The experiments were conducted on the multilingual set MULTEXT (Campione & Veronis 1998). This database contains audio recordings in five European languages (English, French, German, Italian and Spanish) spoken by 50 individuals (5 men and 5 women per language). Each recording represents a 5-sentence text and each speaker read between 10 and 20 of those short texts. Table 1 summarizes the clustered structure of the data, one cluster corresponding to one individual's recordings.

The task is to identify the language spoken on a recording different from those used for learning.

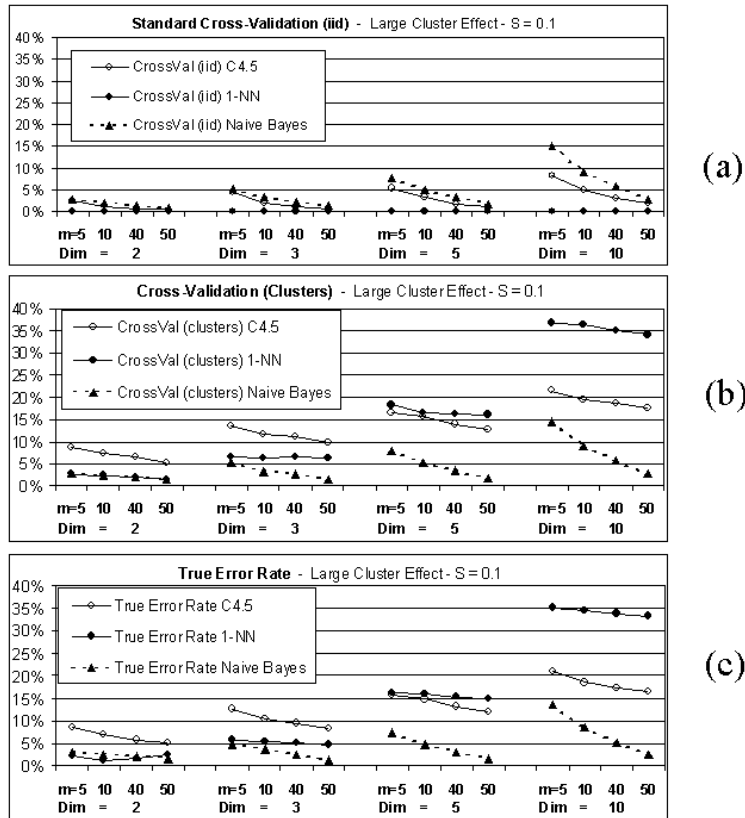


Figure 4: Standard *iid*, cluster cross-validation error rate estimates and true error rates with large cluster effects

4.3 The descriptors

Classical approaches are based on the spectral analysis of the audio signal for which the cluster effect is well-recognized (the spectrum informs on the identity and the language of the speaker). The approach followed here used a parameterized rhythmic space for which the cluster is theoretically less important.

An automatic segmentation of the audio signal in pseudo-syllables was first realized (Farinas & Pellegrino 2001). These units are composed of one or more consonant segments followed by one vocalic segment. They are correlated to the rhythmic structure of the language and can thus be used to identify the language. Each pseudo-syllable is set in a 5-dimensional space:

- Dc (total duration of the pseudo-syllable consonants, in ms);
- Dv (duration of the vocalic segment, in ms);
- Nc (number of consonant segments in the pseudo-syllable, unit free);
- Fo (fundamental frequency of the pseudo-syllable vowel, in Hertz);
- E (relative energy of the vowel, in dB).

For each recording, the parameter means, variances and covariances were computed using all the pseudo-syllables of the sound excerpt. Hence, 20 parameters are available for each statistical individual.

4.4 Comparing various approaches

Many learning algorithms were tested, they have different representation and learning characteristics (Hastie et al. 2001): C4.5 Decision Tree (DT); 1-Nearest Neighbour (1-NN); Naive Bayes (NB); linear

Algorithme	Standard (iid)	Clustered
Decision Tree	25%	35%
1-NN	26%	37%
Naive Bayes	36%	48%
Linear Disc. Analysis	15%	20%
Multi-layer Perceptron	16%	21%
GMM	-	20%

Table 2: Cross-validation error rate estimate for each induction method

discriminant analysis (LDA); multi-layer perceptron (MLP).

In all cases, cross-validation was performed, first not accounting for clustering (different recordings of the same speaker can be used for both learning and testing), and secondly accounting for the clusters (speakers for learning and testing are different).

Finally, a comparison with the EM (Expectation-Maximization) estimation algorithm on Gaussian mixture model (GMM), usually used in speech processing, was performed. Table 2 summarizes the results.

In spite of the small number of characteristics accounted for (average duration of the consonant segments, etc.), rhythmic modeling gives interesting results, in the order of 20% of erroneous identification, whether pattern recognition (GMM) or two usual data mining techniques are used. The complex parameter space seems to handicap DT, 1-NN and NB, more than LDA or MLP.

Moreover, accounting for the clustering modifies the estimated error rate whatever the learning algorithm used. These experiments illustrate that, when working with real data, clustering must be factored in to avoid serious underestimation of the generalization error rate.

5 Related work

The key point of this work is the necessity of accounting for the sample design of the learning dataset when defining the resampling procedure (cross-validation, bootstrap, etc.). Thus, subdividing (learning set, testing set) must be done randomly on the clusters rather than on the individuals; the same is true for leave-one-out.

There is little similar work. Most of the discussion has focussed on the optimal number of parts for cross-validation (Kohavi 1995), on the introduction of sophisticated resampling schemes (Dietterich 1998), or on bias correction with respect to the classifier (Efron & Tibshirani 1997). Some authors have introduced stratified cross-validation, the objective being to maintain the distribution of classes among the subdivisions. There is no sound justification to this, the underlying idea being the reduction of the variability of the models produced at each step. They think that, and the work presented here agrees with their hypothesis, the strategy is only efficient when the initial sample is itself stratified, that is the frequency of each class is explicitly reproduced in the sample (Kohavi 1995).

In this paper, we note that the *iid* cross-validation systematically underestimates the true error rate when we have clustered dataset. It is biased. But, the behavior of the variance of the clustered estimator is not clear. It appears that the standard estimators of the variance of the cross-validation error rate is often underestimated (Bengio & Grandvalet 2004). If we use the same estimator in our context, one can think that we obtain the same result. But it is obvious that it will be necessary to study in detail this assertion which rests only on one intuition.

Another problem is model selection. Some works show that a bad estimation of the generalization error rate can be sufficient for model selection if we obtain a clear picture of the relative performance of the learning algorithms (Petrač 2000). The author claims that using a moderate subsample of the dataset allows to ranking the algorithms. When we want to adapt this approach to clustered dataset, the unit of the sample must be the clusters and not the individuals.

Following the same idea, a bad generalization rate estimation may be sufficient for model selection. One can wonder whether the bias of the standard cross-validation method (*iid* assumption), which underestimates the generalization error rate when instances were sampled from clusters of data, is constant across different learning methods. If it is true, it appears that the *iid* cross-validation can nevertheless used for model selection, whatever the sampling scheme. The answer appears complex, depending on the algorithm characteristics, on the nature of the discrepancy, bias (a systematic difference with respect to the true value) or variance (discrepancies due solely to sampling). Our first results (see section 3.3) show that, using the synthetic data, there is no apparent correlation between the rank of the methods as ordered by standard cross-validation, and their rank as ordered by the cross-validation under clustering. In-depth analysis are still awaited; in the meantime, caution dictates that sampling design be accounted for cross-validation, even if the ultimate goal is model selection.

6 Conclusion

In this paper, it is shown that correct assessment of the predictive model by resampling methods must account for the sampling scheme used for the construction of the learning set. With cluster sampling, stan-

dard cross-validation significantly underestimates the generalization error rate and seems not efficient for model selection.

The approach proposed here can be extended to other sampling strategies (stratification, unequal probability sampling). Then, under those different schemes, the size and direction of the standard cross-validation estimation bias must be determined.

References

- Bay, S. (1999), ‘The UCI KDD archive [<http://kdd.ics.uci.edu/>]’, Irvine, CA: University of California, Department of Computer Science.
- Bengio, Y. & Grandvalet, Y. (2004), ‘No unbiased estimator of the variance of k-fold cross-validation’, *Journal of Machine Learning Research* **5**, 1089–1105.
- Campione, E. & Veronis, J. (1998), A multilingual prosodic database, in ‘Proc. of ICSLP’98’, Sydney.
- Dietterich, T. (1998), ‘Approximate statistical tests for comparing supervised classification learning algorithms’, *Neural Computation* **10**(7), 1895–1924.
- Efron, B. & Tibshirani, R. (1995), Cross-validation and the bootstrap : Estimating the error rate of a prediction rule, Technical Report 176, Department of Statistics, University of Toronto.
- Efron, B. & Tibshirani, R. (1997), ‘Improvements on cross-validation: The 0.632+ bootstrap method’, *JASA* **92**(438), 548–560.
- Farinas, J. & Pellegrino, F. (2001), Automatic rhythm modeling for language identification, in ‘Proc. of Eurospeech ’01’, Aalborg, Scandinavia, pp. 2539–2542.
- Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.
- Kearns, M. J. & Ron, D. (1997), Algorithmic stability and sanity-check bounds for leave-one-out cross-validation, in ‘Computational Learning Theory’, pp. 152–162.
- Kohavi, R. (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, in ‘Proceedings of the International Joint Conference on Artificial Intelligence - IJCAI’95’.
- Lavrac, N. (1999), ‘Selected techniques for data mining in medicine’, *Artificial intelligence in medicine* **16**, 3–23.
- Petrač, J. (2000), Fast subsampling performance estimates for classification algorithm selection, in ‘ECML-00 Workshop on Meta-Learning: Building Automatic Advice Strategies for Model Selection and Method Combination’, pp. 3–14.
- Quinlan, J. (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.
- Shao, J. & Tu, D. (1995), *The Jackknife and Bootstrap*, Springer.
- Stone, M. (1974), ‘Cross-validated choice and assessment of statistical predictions’, *Journal of the Royal Statistical Society B* **36**, 111–147.
- Zisman, M. & Berkling, K. (2001), ‘Automatic language identification’, *Speech Communication* **35**(1).