Proceedings of Statistics Canada symposium 2002
Modelling Survey Data for Social and Economic Research

ERROR RATE ESTIMATE FOR CLUSTER DATA – APPLICATION TO AUTOMATIC SPOKEN LANGUAGE IDENTIFICATION

J.H. Chauchat[1], R. Rakotomalala[1] and F. Pellegrino[2]

ABSTRACT

If the dataset available to machine learning results from cluster sampling, 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. The results are confirmed on a true application of automatic spoken language identification.

KEYWORDS: Data mining, language identification, cross-validation, clustered dataset

1. INTRODUCTION

In data mining, learning is based on a dataset that is not a sample of the universe to which the results are to be generalized.

Measuring the quality of this generalization is called “validation” [STO 74]. This measurement of quality [LAV 99] (generalization error rate, sensitivity, specificity, ROC curve, TQC coefficient [CHA 01], ...) is achieved through re-sampling methods [EFR 95], which are most often applied with the assumption that the dataset used constitutes a realisation of a simple random selection (iid, for unrestricted and identically distributed selections) for the reference universe [DIE 98].

In practice, this hypothesis is seldom validated, the dataset often being the product of cluster samples or (and more generally) two-stage sampling:

- the set of patients in a sample of hospital services,
- the set of students in a sample of classes or schools, or
- the set of “carrots” when drilling for an oil-well sample.

In this article, we show that the results of standard cross-validation (considering that the individuals come from a simple random sample in the initial population) is too optimistic when the dataset comes from a cluster sample. We propose an adaptation of the cross-validation process that takes the sampling method into account. The results highlight the impact of cross-validation on simulated data; we then apply our approach to a true application of automatic spoken language identification.

In section 2, we adapt the cross-validation technique to cluster sampling; in section 3, we use a simulation to compare the distributions of generalization error rate estimates (with and without clusters) and the exact values of this error rate; in section 4, we present the results obtained through a true dataset: it is a question of automatic

[1] Laboratoire ERIC – Université Lyon 2, 5 av. Mendès-France, F69676 Bron
[2] Laboratoire Dynamique Du Langage – UMR 5596 CNRS – Université Lyon 2, 14 av. Berthelot, F-69365 Lyon, cedex 07

recognition of spoken language by a sample of speakers through a physical analysis of their voice audio signal. In section 5, we attempt to situate the method proposed in this article in relation to the different cross-validation variants. Finally, in section 6, we conclude.

2. ADAPTATION OF CROSS-VALIDATION TO CLUSTER SAMPLES

The generalization error rate measures the error propensity of a model based on a learning sample and applied to the whole population. It is seldom measurable since we rarely have access to the whole population, which has to be estimated. In this context, the method of composition of the learning sample is part of the process of constructing the classifier, and it must be evaluated on that basis.

In reality, it often happens that this information is not available, particularly on the basis of test datasets disseminated through Internet servers (UCI for instance [BAY 99]). In fact, the methods and statistics proposed to measure the error rate are based on a simple sampling hypothesis that should be treated with caution.

In this section, we propose an adaptation of standard cross-validation (iid, for unrestricted and identically distributed selections) when the sample is composed of clusters.

2.1. Cross-validation with an iid sample

The usual cross-validation method (in J parts) takes place as follows [STO 74], based on an implicit assumption of an iid sample:

1) *n* individuals extracted from the population;

2) the *n* individuals in the dataset are randomly distributed into *J* parts, respectively sized $n_j = \frac{n}{J}$ individuals;

3) a learning algorithm is applied to the entire base, with the exception of one J part;

4) individuals in part J are subjected to the rules learned in (3), and the T $_j$ error rate is observed for those individuals who were not used in the learning;

5) “generalization error rate” *T* is estimated by $\hat{T} = \frac{1}{J}\sum T_j$.

The $\hat{T}$ estimator is biased: $E(\hat{T}) < T$ since the learning samples used in (2) are size $n\frac{J-1}{J} < n$*;* this bias become negligible when *J* becomes large (with an iid sample, of course); but the random variation of $\hat{T}$ increases and the calculation time increases with J .

2.2. Cross-validation with a cluster sample

If the dataset comes from a cluster sample, then the above standard procedure must be adjusted as follows:

1) n individuals distributed into G clusters (g = 1, . . . , G), respectively sized n_g observations;

2) the G clusters are subdivided into *J* parts, thus the part designated as *j* has $n_j = \sum_{g \in j} n_g$ observations;

3) to 5) the standard procedure is applied.

In general, there is a cluster effect, i.e. the internal variability of the clusters is low compared to the total variability of the universe; thus, for a total size of *n* from the base, the true generalization error rate increases. This must be highlighted in the cross-validation estimation process.

3. APPLICATION ON SIMULATED DATA

The main purpose of using a simulation model is to be able to determine the true error rate. We can either calculate the conceptual error by relying on the distributions used or, based on the data generator, generate as many individuals as wanted in order to make up the test set and thereby estimate the true error rate with controlled precision. This was the approach chosen for this article.

At this point we present an illustrative example of the approach with only two explanatory variables; this makes it easier to interpret the charts.

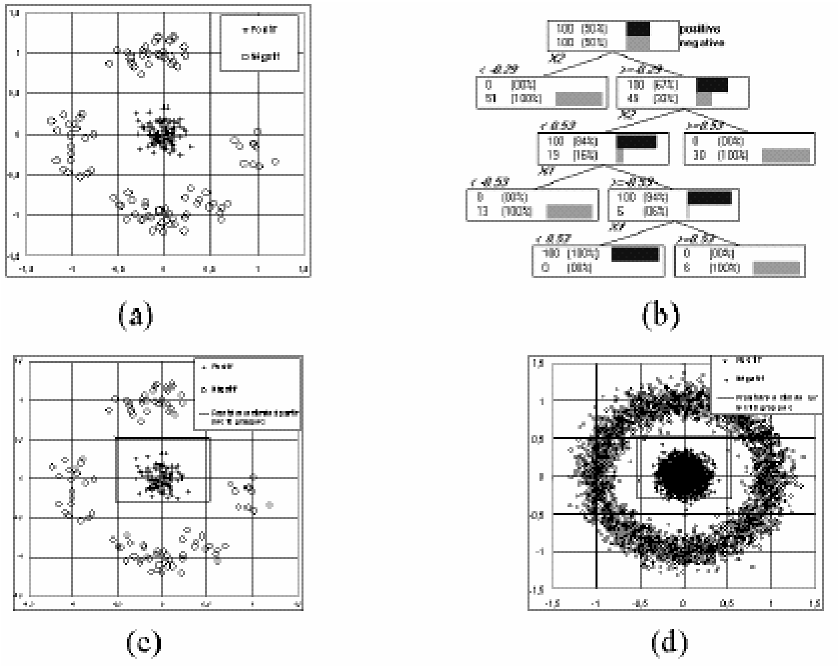
3.1. Model for the simulation

The objective of learning is to distinguish between two classes: the positives (+) and the negatives (o). Individuals in the universe are grouped into classes comprising *m* individuals each, of which $\frac{m}{2}$ are in class "+" and $\frac{m}{2}$ in class "o". Regardless of their cluster, positive individuals in the universe are distributed according to an origin-centred normal bi-dimensional distribution and a variance-covariance matrix s ^{2} × I, where *s* is a constant and *I* is the identity matrix. Negative individuals in a cluster are distributed in the same way but centred on a point on the circle with a 1 radius. These cluster centres (for negative individuals) are random, uniformly distributed on the circle (e.g. figure 1.a). The base consists of g clusters, or $n = m \times$ g individuals.

Hence, there are three parameters in this model: s, the dispersion of each half cluster around its centre, *m*, the number of individuals in each cluster, and *g*, the number of clusters in the learning base.

For each value of these parameters *(s = 0.1 , 0.2 , 0.5 , 1; m = 5, 10, 20, 50;* g = 10), we randomly generated 100 learning bases through this model, as well as a test base of 1000 clusters designed to estimate with reasonable accuracy the true generalization error rate for any learning.
Learning was done through decision trees [QUI 93], which may come close to all borders, whether or not linear, through broken lines comprised of right segments running parallel to one of the axes [ZIG 00].

For each learning sample (Figure 1.a), we have:



(a)

(b)

(c)

(d)

Figure 1. Learning throughout the available sample (a, b, c), application of the grading rule to a large test sample representing the universe (d)

1) construction of the decision tree on the overall sample (Figures 1.b and 1.c) ;
2) calculation of the true error rate for this tree on the large test set (Figure 1.d) ;
3) estimate of the generalization error through cross-validation, taking into account the clusters;
4) estimate of the generalization error through cross-validation without taking it into account (i.e. as if the n individuals were a result of an iid sample).

We return to the operation of the algorithm, based on the example of learning using the base of figure 1.a, which comprises 10 clusters, where $n = G \times m = 10 \times 20 = 200$ individuals. A tree (figure 1.b) is constructed on this basis and the error rate is nil (figure 1.c); the evaluation of the actual error rate on the larger test base is illustrated by figure 1.d.

When cross-validating with J = 10, 1 cluster is eliminated at each stage. One of these stages is illustrated by figure 2: by learning on the 9 clusters at the far left, we get a tree (figure 2.a), which is a poor classification of 6 of the 10 negatives that were set aside (figure 2.b).

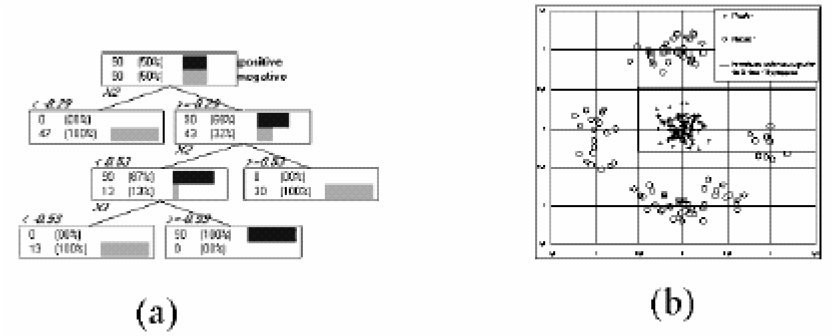


Figure 2. A stage of cross-validation

3.2. Simulation results

The results show a number of elements (Figure 3):

- the actual error rate, Err, increases with s, the relative dispersion of the clusters and decreases with *m*, the size of the clusters;
- standard cross-validation, ignoring the effect of the clusters, greatly underestimates this rate; the relative bias increases with cluster size *m*;
- the error rate estimation bias increases with the cluster effect: it is greatest when ***s*** **=** 0 0, which is to say when all individuals in a cluster are identical;
- cross-validation that takes into account the cluster effect slightly underestimates the actual error rate; this was expected since, as indicated above, cross-validation is slightly biased by the fact that it uses, at each stage, a fraction of the sample available for constructing the forecast model.

4. APPLICATION TO ACTUAL SPOKEN LANGUAGE IDENTIFICATION DATA

4.1. Issues addressed

Identification of spoken languages based on recordings is a new area of spoken language identification. At a time of communications media globalization, there are a number of issues, both in Human-Machine Interfaces and human speech assistance.

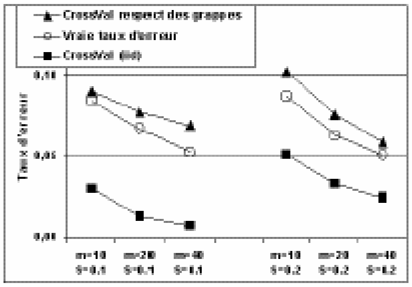


Figure 3. *Actual and estimated error rate for G=10 clusters: s = 0.1, s = 0.2 and m = 10, 20 and 40. Average of 100 simulations for each case.*

Most of the approaches developed to date are based on statistical modeling of the phonetic (nature of sounds) and didactic (sound linkage) features of the different languages addressed [ZIS 01]. Such approaches call for the use of large volumes of recordings accompanied by their phonetic transcripts (fully supervised learning).

Data mining, when associated with the recognition of innovating parameters, can lead to convincing results while requiring only partly supervised learning and a smaller number of learning data.

4.2. Task description and data used

The experiences are based on the multilingual corpus, MULTEXT [CAM 98]. This dataset contains recordings from 5 European languages (German, English, Spanish, French and Italian) spoken by 150 speakers (5 men and 5 women per language). Each recording corresponds to the reading of a text comprising approximately 5 sentences, and each speaker reads between 10 and 20 of these texts. Table 1 summarizes the data cluster structure, with one cluster corresponding to the set of recordings produced by a speaker.

The task involves the identification of the language spoken in a recording separate from those used for learning.

4.3. Descriptor space

Unlike the classical approaches, which are based on the spectral information contained in the signal, for which the cluster effect is well known, (the spectrum provides just as much information about the speaker's identity as about the language spoken), we find ourselves in a rhythmic parametric space where the cluster effect is theoretically less evident.

Language	Speakers	Recordings / Speaker	Avg. Durat. / Recording
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. Cluster structure of the MULTEXT corpus.

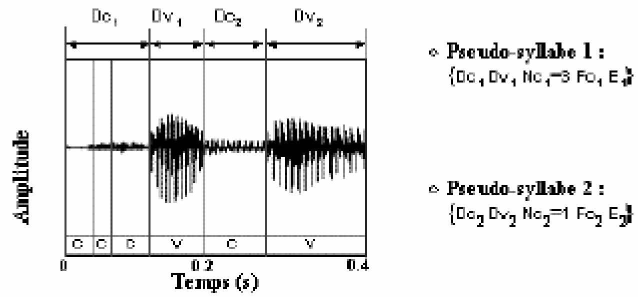


Figure 4. Example of parametrization into pseudo-syllables. The extract is segmented into consonant segments (C) and vowel segments (V), which form two pseudo-syllables called 1 and 2. Each pseudo-syllable is parametrized by 5 parameters (Fo and E are not represented).

Starting with the acoustic signal, we do an automatic segmentation into pseudo-syllables [FAR 01]. These units comprise one or more consonantic segments followed by a vocalic segment (Figure 4). They are correlated with the rhythmic structure of the language and can therefore help recognize languages. Each pseudo-syllable is characterized by 5 parameters:

- Dc (total duration of the pseudo-syllable consonants, in ms);
- Dv (duration of the vocalic segment, in ms);
- Nc (number of consonantic segments of the pseudo-syllable, without identity) ;
- Fo (fundamental frequency of the vowel in the syllable, in Hertz) ;
- E (relative energy of the vowel, in dB).

Algorithm	**Standard (iid)**	**In clusters**
Decision tree	25%	35%
Discriminant analysis	15%	20%
Multilayer perceptron	16%	21%
GMM	-	20%

Table 2. *Error rate (in cross-validation) for the different approaches studied*

For each recording, we calculated the averages, variances and co-variances for these parameters on the set of pseudo-syllables in the passage. Thus, we had a set of 20 parameters for each statistical individual (recording).

4.4. Comparison between the different approaches

Several learning methods were tested:

- decision tree (DT) [QUI 93];
- linear discriminant analysis (LDA) [FIS 36] ;
- multilayer perceptron (MLP) [MIT 97].

In every case, there was a cross-validation, on the one hand without taking into account the cluster effect (different recordings, while from a single speaker, can be used for learning and testing), on the other hand by taking into account this effect (all learning and test speakers are distinct).

Finally, there was a comparison with a modelization of the data through a blend of Gaussian laws (GMM) estimated with the EM (ExpectationMaximization) algorithm, a classically used approach in spoken language identification [REY 95].

Table 2 summarizes the results.

4.5. Discussion

Despite the small number of characteristics taken into account (average of vocalic and consonantic durations, etc.), the rhythmic modelization approach produced some very interesting results, in the order of 20% of false identifications, whether through a form recognition approach (GMM) or knowledge extraction algorithms. In this latter case, the complexity of the parameter space seems to penalize the decision tree algorithm in relation to the LDA and MLP approaches.

Moreover, taking into account the appurtenance of individuals to clusters significantly changes the error rate obtained: these tests confirm that, when working on actual data, it is essential to take this effect into account, given the danger of significantly underestimating the actual error rates (when the rules are applied to the entire universe).

5. DISCUSSIONS AND SIMILAR WORKS

The main principle that should be extracted from our approach is that it is necessary to take into account the means of constituting the learning sample when defining the evaluation procedure through re-sampling (cross-validation, bootstrap [EFR 95], etc.). Thus, in the subdivision diagrams (learning set – test set), we have to do a random selection, not on the statistical individuals, but on the clusters; the same applies for the jacknife (leave-one-out) [WEI 91], the subdivisions have to be applied to the clusters.

There are not many works similar to ours. Discussion has mainly focused on the number of parts in the cross-validation [KOH 95], the introduction of evolved re-sampling diagrams [DIE 98], or the correction of the bias, taking into account the characteristics of the classifier [TIB 96]. Nonetheless, since the works of Breiman et al. [BRE 84], a number of works have introduced the cross-validation stratified re-sampling plan. The objective is to respect the distribution of classes in each subdivision. There is no true justification for this, the underlying idea being to reduce the variability of the models produced with each passage. Nonetheless, some authors [KOH 95] believe, and in this regard agree with our point of view, that this strategy is only truly effective if the initial sample was extracted from the population in a stratified manner, which is to say that the probabilities of occurrence of each class during the construction of the sample was explicitly respected.

Finally, while we now know that a standard cross-validation (iid) underestimates the error rate when the sample is made of clusters, we still do not know whether the deviation from the actual error rate is the same regardless of the method used. This point is important if we want to select the best model in terms of the cross-validation-calculated error rate from among all the models constructed [STO 74]. The answer appears to be quite complex, as it depends on the characteristics of the algorithm used and the nature of the deviation, which may be due to the bias (a systematic deviation in terms of the actual error value) or the variance (the variability due to the sample) [KOH 95]. Without specific answers to these questions, caution dictates that we respect the sampling plan when constructing the cross-validation even when the objective is model selection.

6. CONCLUSION

This article demonstrates that the method of composition of the learning set has to be taken into account if the prediction model is to be accurately evaluated through re-sampling methods. With clusters, normal cross-validation, where individuals are assumed to come from a simple random selection from the initial population, significantly underestimates the actual error rate.

Our approach may be extended to the different sampling methods (stratification, with unequal probabilities). All that remains for the different cases is to determine which direction and to what extent lack of knowledge about the method of constitution may influence the estimation of the quality of the model resulting from the standard cross-validation.

REFERENCES

[BAY 99] BAY S., “The UCI KDD Archive [http ://kdd.ics.uci.edu]”, Irvine, CA : University of California, Department of Computer Science, 1999.

[BRE 84] BREIMAN L., FRIEDMAN J., OLSHEN R., STONE C., *Classification and Regression Trees,* California : Wadsworth International, 1984.

[CAM 98] CAMPIONE E., VERONIS J., “A multilingual prosodic database”, *Proc. of* ICSLP’98, Sidney, 1998.

[CHA 01] CHAUCHAT J., RAKOTOMALALA R., PELLETIER C., CARLOZ M., “TQC : un indice d’évaluation de la détection d’évènements rares - Une application au ciblage en marketing", *Extraction et gestion des connaissances, vol. 1,* no 1-2, 2001, p. 155-160.

[DIE 98] DIETTERICH T., “Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms”, *Neural Computation, vol. 10,* no 7, 1998, p. 1895–1924.

[EFR 95] EFRON B., TIBSHIRANI R., “Cross-validation and the Bootstrap: Estimating the Error rate of a Prediction rule”, rapport no 176, 1995, Department of Statistics, University of Toronto.

[FAR 01] FARINAS J., PELLEGRINO F., “Automatic Rhythm Modeling for Language Identification", *Proc. of Eurospeech ’01,* Aalborg, Scandinavia, September 2001, p. 2539-2542.

[FIS 36] FISHER R., “The use of multiple measurements in taxonomic problems”, *Annals of Eugenics, vol.* 7, 1936, p. 179–188.

[KOH 95] KOHAVI R., “A study of cross-validation and bootstrap for accuracy estimation and model selection”, *Proceedings of the International Joint Conference on Artificial Intelligence* - IJCAI’95, 1995.

[LAV 99] LAVRAC N., “Selected techniques for data mining in medicine”, *Artificial intelligence in medicine, vol.* 16, 1999, p. 3-23.

[MIT 97] MITCHELL *T., Machine learning,* McGraw Hill, 1997.

[QUI 93] QUINLAN J., C4.5 : *Programs for Machine Learning,* Morgan Kaufmann, San Mateo, CA, 1993.

[REY 95] REYNOLDS D., “Speaker Identification and Verification using Gaussian Mixture Speaker Models”, *Speech Communication, vol.* 17, no 1-2, 1995, p. 91-108.

[STO 74] STONE M., “Cross-validatory choice and assessment of statistical predictions”, *Journal of the Royal Statistical Society, vol.* B 36, 1974, p. 111-147.

[TIB 96] TIBSHIRANI R., “Bias, variance and prediction error for classification rules”, rapport, 1996, Department of preventive Medecine and Biostatistics and Department of Statistics, University of Toronto.

[WEI 91] WEISS S., KULIKOWSKI C., *Computer Systems that Learn,* Morgan Kaufmann, San Mateo, CA, 1991.

[ZIG 00] ZIGHED D., RAKOTOMALALA R., *Graphes d’Induction - Apprentissage et Data Mining,* Hermes, 2000.

[ZIS 01] ZISMAN M., BERKLING K., “Automatic language identification”, *Speech Communication, vol.* 35, no 1, 2001.