Recueil du Symposium 2002 de Statistique Canada
Modélisation des données d'enquête pour la recherche économique et sociale

ESTIMATION DU TAUX D'ERREUR SUR DONNEES EN GRAPPES - APPLICATION À LA RECONNAISSANCE DE LA PAROLE

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

RÉSUMÉ

Lorsque l'ensemble de données est issu d'un échantillonnage par grappes, les résultats de la validation croisée standard (considérant que les individus sont issus d'un tirage aléatoire simple dans la population initiale) est trop optimiste, sous-estimant significativement le vrai taux d'erreur. Nous proposons dans cet article une modification du procédé de validation croisée qui tient compte du mode d'échantillonnage. Nous étayons nos résultats en montrant, dans un premier temps, le comportement de la validation croisée sur données simulées ; puis, dans un second temps, nous appliquons notre approche sur un problème de reconnaissance de la parole.

MOTS-CLÉS : Extraction de connaissances à partir de données, Reconnaissance de la parole, Validation croisée, données en grappes, mesure de la qualité de la connaissance

1. INTRODUCTION

En extraction de connaissances à partir de données, l'apprentissage est réalisé sur une base de données qui n'est qu'un échantillon de l'univers auquel les résultats doivent être généralisés.

La mesure de la qualité de cette généralisation est nommée "validation" [STO 74]. Cette mesure de la qualité [LAV 99] (taux d'erreur en généralisation, sensibilité, spécificité, courbe ROC, coefficient TQC [CHA 01], ...) est obtenue par des méthodes de rééchantillonnage [EFR 95] qui, le plus souvent, sont appliquées avec l'hypothèse que la base de donnée utilisée constitue une réalisation d'un échantillon aléatoire simple (iid, pour tirages indépendants et identiquement distribués) de l'univers de référence [DIE 98].

En pratique, cette hypothèse est rarement réalisée, la base est souvent le produit de tirages en "grappes", ou bien (plus généralement) de tirages à deux degrés:

- l'ensemble des patients d'un échantillon de services hospitaliers,
- l'ensemble des élèves d'un échantillon de classes ou d'écoles,
- l'ensemble des "carottes" lors du forage d'un échantillon de puits de pétrole.

Dans cet article, nous montrons que les résultats de la validation croisée standard (considérant que les individus sont issus d'un tirage aléatoire simple dans la population initiale) est trop optimiste lorsque l'ensemble de données est issu d'un échantillonnage par grappes. Nous proposons alors une modification du procédé de validation croisée qui tient compte du mode d'échantillonnage. Nous étayons nos résultats en montrant, dans un premier temps, le comportement de la validation croisée sur données simulées ; puis, dans un second temps, nous appliquons notre approche sur un problème de reconnaissance de la parole.

Dans la section 2, nous adaptons la technique de validation croisée au cas des tirages en grappes ; dans la section 3, nous utiliserons un exemple simulé pour comparer les distributions des estimations du taux d'erreur en généralisation (en utilisant, ou non, les grappes) et les valeurs exactes de ce taux d'erreur ; dans la section 4, nous présenterons les résultats obtenus sur une base de données réelle : il s'agit de reconnaître automatiquement la langue parlée par un échantillon de locuteurs à partir de l'analyse physique du signal audio de leurs voix. Dans la cinquième section, nous essayerons de situer la méthode proposée dans cet article par rapport aux différentes variantes de la

[1] Laboratoire ERIC – Université Lyon 2, 5 av. Mendès-France, F-69676 Bron
[2] Laboratoire Dynamique Du Langage - UMR 5596 CNRS - Université Lyon 2, 14 av. Berthelot, F-69365 Lyon cedex 07

validation croisée. Enfin, nous conclurons dans une sixième et dernière section.

2. ADAPTATION DE LA VALIDATION CROISEE AUX ECHANTILLONS EN GRAPPES

Le taux d'erreur en généralisation mesure la propension à l'erreur du modèle, construit à l'aide de l'échantillon d'apprentissage, appliqué sur toute la population. Il est rarement mesurable parce que nous avons très rarement accès à la totalité de la population, nous devons donc l'estimer. Dans ce cadre, le mode de constitution de l'échantillon d'apprentissage fait partie du processus de construction du classificateur, son évaluation doit en tenir compte.

En réalité, il apparaît très souvent que cette information n'est pas disponible, notamment sur les bases de données tests diffusés sur les serveurs internet (UCI par exemple [BAY 99]). De fait, les méthodes et statistiques proposées pour mesurer le taux d'erreur reposent sur une hypothèse d'échantillonnage simple sujette à caution.

Dans cette section, nous proposons une adaptation de la validation croisée standard (iid, pour tirages indépendants et identiquement distribués) lorsque l'échantillon est composé de grappes.

2.1. Validation croisée avec un échantillon iid

La méthode de validation croisée (en J parties) usuelle procède comme suit [STO 74], supposant implicitement un échantillon iid:

1) on dispose de *n* individus extraits de la population ;
2) les *n* individus de la base sont répartis aléatoirement en *J* parties, de taille respective $n_j = \frac{n}{J}$ individus ;
3) on met en oeuvre l'algorithme d'apprentissage sur toute la base sauf une partie j ;
4) on applique, sur les individus de la partie *j* mise de côté, les règles apprises en (3), et on observe le taux d'erreur T_j sur ces individus qui n'ont pas servi à l'apprentissage ;
5) le "taux d'erreur en généralisation" *T* est estimé par $\hat{T} = \frac{1}{J}\sum T_j$.

L'estimateur $\hat{T}$ est biaisé : $E(\hat{T}) < T$ car les échantillons d'apprentissage utilisés en (2) sont de taille $n\frac{J-1}{J} < n$; ce biais devient négligeable quand *J* devient grand (si l'échantillon est iid, bien sûr) ; mais la variation aléatoire de $\hat{T}$ augmente, et le temps de calcul croît comme J .

2.2. Validation croisée avec un échantillon en grappes

Si la base provient d'un échantillon "en grappes", alors la procédure standard ci-dessus doit être modifiée de la manière suivante:

1) on dispose de n individus, repartis en G grappes (g = 1, . . . , G) de taille respective n_g observations;
2) les G grappes sont subdivisés en *J* parties, la partie numéro *j* comporte donc $n_j = \sum_{g \in j} n_g$ observations;

3 à 5) nous appliquons la procédure standard.

En général, il y a un "effet de grappe", c'est à dire que la variabilité interne aux grappes est faible comparée à la variabilité totale de l'univers ; alors, pour une taille totale *n* donnée de la base, le vrai taux d'erreur en généralisation augmente. Ceci doit être mis en évidence dans le procédé d'estimation par validation croisée.

3. APPLICATION SUR DONNÉES SIMULÉES

Le principal intérêt d'utiliser un modèle de simulation est que l'on a la possibilité de connaître le vrai taux d'erreur. En effet, nous pouvons : soit calculer l'erreur théorique en nous appuyant sur les distributions utilisées ; soit, en disposant du générateur de données, générer autant d'individus que l'on veut afin de constituer l'ensemble test et estimer ainsi le vrai taux d'erreur avec une précision que l'on contrôle. C'est le choix que nous avons fait dans cet article.

Nous présentons ici un exemple illustratif de la démarche avec deux variables explicatives seulement ; les graphiques sont ainsi plus faciles à interpréter.

3.1. Le modèle pour la simulation

L'objectif de l'apprentissage sera de distinguer deux classes : les positifs (+) et les négatifs (o). Dans l'univers les individus sont regroupés en classes comportant chacune *m* individus, dont $\frac{m}{2}$ sont dans la classe "+" et $\frac{m}{2}$ dans la classe "o". Dans l'univers, les individus positifs sont distribués, quelque soit leur grappes, selon une loi normale bi-dimensionnelle centrée à l'origine et de matrice de variance-covariance $s^2 \times I$, où *s* est une constante et *I* est la matrice unité. Les individus négatifs d'une grappe sont distribués selon la même loi, mais centrée sur un point du cercle de rayon 1. Ces centres de grappes (pour les individus négatifs) sont aléatoires, de distribution uniforme sur le cercle (par exemple figure 1.a). La base est composée de g grappes, soit $n = m \times g$ individus.

Il y a donc trois paramètres dans ce modèle : s , la dispersion de chaque demigrappe autour de son centre, *m*, le nombre d'individus dans chaque grappe, et *g*, le nombre de grappes dans la base d'apprentissage.

Pour chaque valeur de ces paramètres *(s = 0.1, 0.2, 0.5, 1; m = 5, 10 , 20 , 50;* g = 10), on a généré aléatoirement 100 bases d'apprentissage selon ce modèle, ainsi qu'une base test de 1, 000 grappes destinée à estimer assez précisément le vrai taux d'erreur en généralisation de tout apprentissage.

L'apprentissage a été réalisé par les arbres de décision [QUI 93], qui peuvent approcher toutes frontières, linéaires ou non, par des lignes brisées composées de segments de droites parallèles à l'un ou l'autre des axes [ZIG 00].

Pour chaque échantillon d'apprentissage (Figure 1.a), on a:

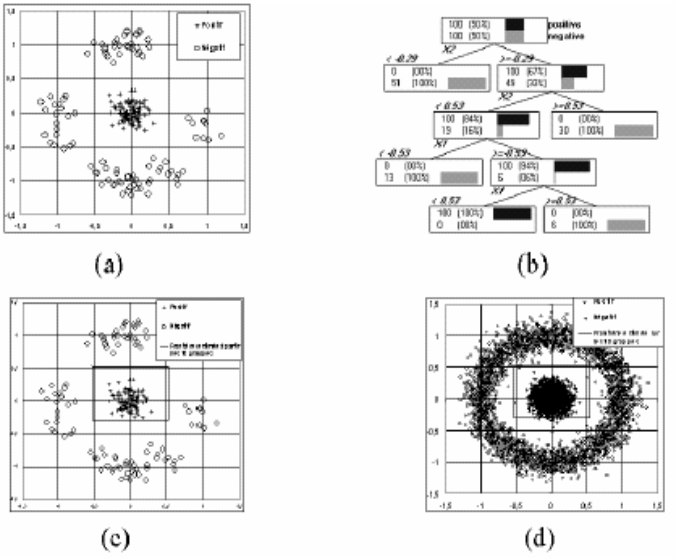

Figure *1.* *Apprentissage sur l'ensemble de l'échantillon disponible (a, b, c), application de la règle de classement sur un gros échantillon test représentant l'univers (d)*

1) construit l'arbre de décision sur l'ensemble de l'échantillon (Figures 1.b et 1.c) ;
2) calculé le vrai taux d'erreur de cet arbre sur la grande base-test (Figure 1.d) ;
3) estimé l'erreur en généralisation par validation croisée en tenant compte des grappes ;
4) estimé l'erreur en généralisation par validation croisée sans en tenir compte (c'est à dire en faisant comme si les n individus résultaient d'un tirage iid).

Reprenons et commentons le fonctionnement de l'algorithme en nous appuyant sur l'exemple d'un apprentissage utilisant la base de la figure 1.a qui comprend 10 grappes, soit $n = G \times m = 10 \times 20 = 200$ individus. Un arbre (figure 1.b) est construit sur cette base et le taux d'erreur est nul (figure 1.c) ; l'évaluation du taux d'erreur réel, sur la grande base test, est illustré par la figure 1.d.

Lors de la validation croisé avec J = 10, on élimine 1 grappe à chaque étape. L'une de ces étapes est illustré par la figure 2 : en apprenant sur les 9 grappes les plus

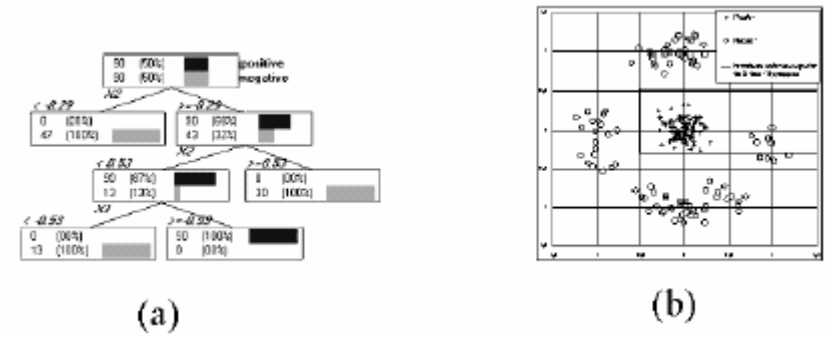

Figure 2. *Une étape de la validation croisée*

à gauche, on obtient un arbre (figure 2.a) qui classe mal 6 des 10 négatifs mis de coté (figure 2.b).

3.2. Résultats de la simulation

Les résultats montrent plusieurs éléments (Figure 3):

– le vrai taux d'erreur Err augmente avec s, la dispersion relative des grappes, et décroît avec *m*, la taille des grappes;
– la validation croisée standard, ignorant l'effet de grappes, sous-estime fortement ce taux ; le biais relatif augmente avec la taille *m* des grappes;
– le biais d'estimation du taux d'erreur croît avec l'effet de grappes : ce dernier est maximum quand $\boldsymbol{s} = 0$, c'est à dire quand tous les individus d'une grappe sont identiques;
– la validation croisée tenant compte de l'effet de grappe sur-estime légèrement le vrai taux d'erreur ; ceci était attendu car comme nous l'annoncions plus haut, la validation croisée est légèrement biaisée du fait qu'elle utilise, à chaque étape, une fraction de l'échantillon disponible pour construire le modèle de prédiction.

4. APPLICATION A DES DONNEES REELLES ENTRAITEMENT DE LA PAROLE

4.1. Problématique abordée

L'identification des langues à partir d'enregistrements sonores est un domaine récent du traitement automatique de la parole. A une époque où les médias de communication s'internationalisent, les enjeux sont nombreux, tant dans le domaine des Interfaces Homme-Machine que dans l'assistance au dialogue humain.

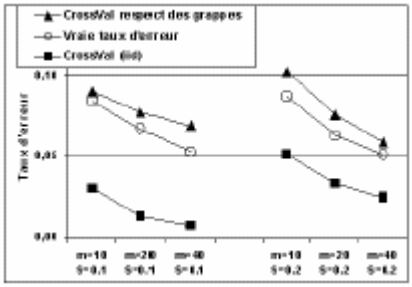

Figure 3. *Taux d'erreur correct et estimés pour G=10 grappes : s = 0.1, s = 0.2 et m = 10, 20 et 40. Moyennes sur 100 simulations pour chaque cas.*

La plupart des approches développées jusqu'à présent se basent sur une modélisation statistique des caractéristiques phonétiques (nature des sons) et phonotactiques (règles d'enchaînement des sons) des différentes langues traitées [ZIS 01]. De telles approches nécessitent de disposer de grandes quantités d'enregistrements accompagnés de leurs transcriptions phonétiques (apprentissage entièrement supervisé).
Les techniques d'extraction des connaissances, associées à une prise en compte de paramètres innovants, peuvent permettre d'obtenir des résultats convaincants tout en ne requérant qu'un apprentissage partiellement supervisé et un nombre de données d'apprentissage réduit.

4.2. Description de la tâche et des données utilisées

Les expériences sont réalisées à partir du corpus multilingue MULTEXT [CAM 98]. Cette base de données contient des enregistrements issus de 5 langues européennes (allemand, anglais, espagnol, français et italien) prononcés par 50 locuteurs (5 hommes et 5 femmes par langue). Chaque enregistrement correspond à la lecture d'un texte d'environ 5 phrases, et chaque locuteur a prononcé entre 10 et 20 de ces textes. Le Tableau 1 résume la structure en grappe des données, une grappe correspondant à l'ensemble des enregistrements produits par un locuteur.

La tâche consiste à identifier la langue parlée dans un enregistrement distinct des enregistrements utilisés pour l'apprentissage.

4.3. L'espace des descripteurs

Contrairement aux approches classiques, basées sur l'information spectrale contenue dans le signal, pour lesquelles l'effet de grappe est bien connu (le spectre renseignant tout autant sur l'identité du locuteur que sur la langue parlée), nous nous sommes placés dans un espace paramétrique rythmique, pour lequel l'effet de grappe est théoriquement moins marqué.

Langue	Locuteurs	Enregistr. / Locuteur	Durée moy. / Enregistrement
Allemand	10	20	21,9
Anglais	10	15	17,6
Espagnol	10	15	20,9
Français	10	10	21,9
Italien	10	15	21,7
TOTAL	50	750	

Tableau 1. Structure en grappes du corpus MULTEXT.

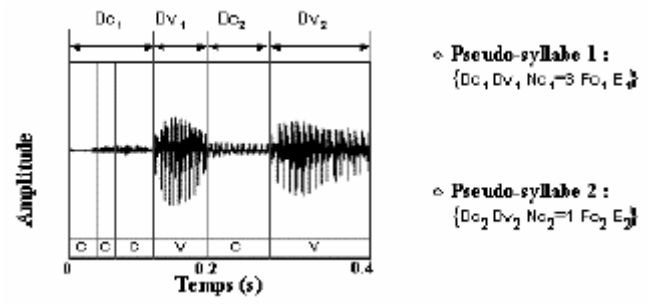

Figure 4. *Exemple de paramétrisation en pseudo-syllabes. L'extrait est segmenté en segments consonnes (C) et segments voyelles (V) qui forment deux pseudo-syllabes notées 1 et 2. Chaque pseudo-syllabe est paramétrée par 5 paramètres (Fo et E ne sont pas représentées).*

A partir du signal acoustique, une segmentation sous forme de pseudo-syllabes est réalisée de manière automatique [FAR 01]. Ces unités se composent d'un ou plusieurs segments consonantiques suivi(s) d'un segment vocalique (Figure 4). Elles sont corrélées à la structure rythmique de la langue et peuvent donc permettre de reconnaître les langues. Chaque pseudo-syllabe est caractérisée par 5 paramètres:

- Dc (durée totale des consonnes de la pseudo-syllabe, en ms) ;
- Dv (durée du segment vocalique, en ms) ;
- Nc (nombre de segments consonantiques de la pseudo-syllabe, sans unité) ;
- Fo (fréquence fondamentale de la voyelle de la syllabe, en Hertz) ;
- E (énergie relative de la voyelle, en dB).

Algorithme	**Standard (iid)**	**En grappes**
Arbre de décision	25%	35%
Analyse discriminante	15%	20%
Perceptron multicouches	16%	21%
GMM	-	20%

Tableau 2. *Taux d'erreur (en validation croisée) pour les différentes approches étudiées*

Pour chaque enregistrement, on a calculé les moyennes, variances et covariances de ces paramètres sur l'ensemble des pseudo-syllabes du passage. On dispose donc d'un ensemble de 20 paramètres pour chaque individu statistique (enregistrement).

4.4. Comparaison des différentes approches

Plusieurs méthodes d'apprentissage ont été testées:
- arbre de décision (DT) [QUI 93] ;
- analyse linéaire discriminante (LDA) [FIS 36] ;
- perceptron multicouche (MLP) [MIT 97].

Dans tout les cas, une validation croisée a été effectuée, d'une part sans tenir compte de l'effet de grappes (des enregistrements différents, mais issus d'un même locuteur peuvent être utilisés pour l'apprentissage et le test) et d'autre part en prenant cet effet en compte (les ensembles de locuteurs d'apprentissage et de tests sont distincts).

Enfin, une comparaison a été faite avec une modélisation des données par un mélange de lois gaussiennes (GMM) estimées avec l'algorithme EM (ExpectationMaximization), approche classiquement employée en traitement de la parole [REY 95].

Le tableau 2 résume les résultats obtenus.

4.5. Discussion

Malgré le petit nombre de caractéristiques prises en compte (moyenne des durées vocaliques et consonantiques, etc.), l'approche par modélisation rythmique permet d'obtenir des résultats très intéressants, de l'ordre de 20 % de mauvaise identification, que ce soit par une approche issue de la reconnaissance des formes (GMM) ou par des algorithmes d'extraction des connaissances. Dans ce dernier cas, la complexité de l'espace des paramètres semble pénaliser l'algorithme des arbres de décision par rapport aux approches LDA et MLP.

De plus, la prise en compte de l'appartenance des individus à des grappes modifie de manière significative le taux d'erreur obtenu : ces expériences confirment que, lorsque l'on travaille sur des données réelles, la prise en compte de cet effet est indispensable, sous peine de sous estimer de manière importante les taux d'erreurs réels (lorsqu'on appliquera les règles à l'ensemble de l'univers).

5. DISCUSSIONS ET TRAVAUX SIMILAIRES

Le principe fort qu'il faut retenir de notre approche est qu'il est nécessaire de tenir compte du mode de constitution de l'échantillon d'apprentissage lors de la définition de la procédure d'évaluation par ré-échantillonnage (validation croisée, bootstrap [EFR 95], etc.). Ainsi, dans les schémas de subdivision (ensemble apprentissage - ensemble test), on doit effectuer une partition aléatoire, non pas sur les individus statistiques, mais sur les grappes ; il en est de même pour le jacknife (leave-one-out) [WEI 91], les subdivisions doivent être réalisées sur les grappes.

Il existe peu de travaux similaires au notre. Le débat a surtout porté sur le nombre optimal de parties dans la validation croisée [KOH 95], l'introduction de schémas de ré-échantillonnage évolués [DIE 98], ou la correction du biais tenant compte des caractéristiques du classifieur [TIB 96]. Néanmoins, depuis les travaux de Breiman et al. [BRE 84], plusieurs travaux ont introduit le schéma de ré-échantillonnage stratifié en validation croisée. L'objectif est de respecter la répartition des classes dans chaque subdivision. Il n'y a pas de véritable justification à cela, l'idée sous-jacente est de réduire la variabilité des modèles produits lors de chaque passage. Néanmoins, certains auteurs [KOH 95] pensent, et en ce sens rejoignent notre point de vue, que cette stratégie n'est véritablement efficace que si l'échantillon initial a été extrait de manière stratifiée dans la population, c'est-à-dire on a respecté de manière explicite les probabilités d'occurrence de chaque classe lors de la construction de l'échantillon.

Enfin, si nous savons maintenant qu'une validation croisée standard (iid) sousestime le taux d'erreur lorsque l'échantillon est constitué de grappes, nous ne savons pas en revanche si l'écart avec le vrai taux d'erreur est le même quelle que soit la méthode utilisée. Ce point est d'importance si nous voulons sélectionner, parmi un ensemble de modèles construits, celle qui se révèle être la meilleure au sens du taux d'erreur calculé en validation croisée [STO 74]. La réponse semble assez complexe, elle dépend à la fois des caractéristiques de l'algorithme utilisé et de la nature de l'écart, qui peut être dû au biais (un écart systématique par rapport à la vraie valeur de l'erreur) ou à la variance (la variabilité dûe à l'échantillon) [KOH 95]. En l'absence de réponses précises à ces

questions, la prudence impose que l'on respecte le schéma d'échantillonnage lors de la construction de la validation croisée même lorsque l'objectif est la sélection de modèle.

6. CONCLUSION

Cet article montre que l'on doit tenir compte du mode de constitution de l'ensemble d'apprentissage si l'on veut évaluer correctement le modèle prédictif à l'aide des méthodes de ré-échantillonnage. Dans le cas des grappes, la validation croisée standard, considérant que les individus sont issus d'un tirage aléatoire simple de la population initiale, sous-estime de manière significative le vrai taux d'erreur.

Notre approche peut être étendue aux différents modes d'échantillonnage (stratification, à probabilités inégales). Reste à savoir dans les différents cas, dans quel sens et dans quelle mesure l'ignorance du mode de constitution peut influer sur l'estimation de la qualité du modèle issue de la validation croisée standard.

RÉFÉRENCES

[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.