# Critères d'information pour la sélection de variables

Mohamed Abadi[1], Olivier Alata[1], Christian Olivier[1], Enguerran Grandchamp[2] et Majdi Khoudeir[1]

[1] Université de Poitiers,
Institut XLIM-SIC, UMR CNRS 6172,
BP. 30179, 86962 Futuroscope-Chasseneuil Cedex, France.
{abadi,alata,olivier,khoudeir}@sic.univ-poitiers.fr

[2] Université des Antilles et de la Guyane,
Laboratoire LAMIA,
Campus de Fouillole, 97157, Pointe-à-Pitre, Guadeloupe, France
egrandch@univ-ag.fr

**Résumé** Cet article étudie les potentialités des critères d'information pour la sélection des variables. Ces critères sont intégrés dans des schémas de parcours de sous-ensembles candidats afin de sélectionner ceux qui les minimisent. Le choix d'un sous-ensemble dépend de la qualité de l'approximation des densités de probabilité de ces variables. A cet effet, nous avons adopté une approche exploitant les histogrammes optimisés à l'aide des principes en lien avec le codage arithmétique adaptatif. Des tests sur des données simulées et de référence sont réalisés. Ils montrent par le biais des taux de bonne classification, déterminés en appliquant plusieurs algorithmes de classification, l'importance d'un tel outil et sa capacité à choisir les sous-ensembles qui caractérisent mieux les classes vis-à-vis de données.

**Mots clés** Critères d'information, sélection de variables, estimation et sélection des histogrammes, schémas de parcours de variables.

**Abstract** This paper introduces the information criteria in feature selection framework. Information criteria are integrated in feature selection scheme to select subset candidates. The accuracy of the proposed approach is based on the quality of the probability density approximations of these features. They are obtained using histograms optimized thanks to the adaptive arithmetic coding principles. Tests on simulated data and references are made. Multiple classifiers are used. The correct classification rate shows, the importance of this tools and its ability to select a best subsets. Generally, the subsets produce a good characterization of classes which the data belong.

**Key words** Information criteria, feature selection, histogram estimation and selection, feature scheme search.

## 1 Introduction

La sélection de variables connue aussi sous le nom de sélection d'attributs est considérée comme l'une des étapes clés dans plusieurs domaines (ex. l'apprentissage automatique, la modélisation de données,

le data-mining, etc. [1, 2, 3]). Elle permet de réduire la dimensionnalité des données en identifiant un ou plusieurs sous-ensembles appartenant à l'ensemble d'origine qui est composé par toutes les variables. Par exemple soient $X = \{x_i, i = 1, \cdots, M\}$ une observation à $M$ variables qui représente la classe $k$. L'objectif est donc de trouver parmi les $M$ variables les $m$ variables ($m < M$) qui caractérisent au mieux $k$. Dans ce cas, le nombre total de sous-ensembles qui peuvent être construit à partir de $M$ variables est $2^M - 1$ et ceux dont la dimension est inférieure ou égale à $m$ est $\sum_{i=1}^{m} \binom{M}{i}$. Plusieurs algorithmes ont été développés pour construire les sous-ensembles car une recherche exhaustive sur les $M$ variables est très couteuse en temps de calcul. Pour plus de détails, le lecteur intéressé peut se référer à [4]. Nous nous intéressons dans cette étude aux algorithmes basés sur des approches séquentielles et plus particulièrement aux schémas de sélection progressive (SFS pour Sequential Forward Selection) et rétrograde (SBS pour Séquential Backward Selection). L'avantage est qu'ils sont faciles à implémenter, indépendants des algorithmes de classification et des critères de décision. Donc ces schémas offrent la possibilité de comparer aisément plusieurs critères entre eux. L'important est de trouver les sous-ensembles qui traduisent au mieux l'information recherchée. Ceci met en avant l'importance du choix du critère de sélection.

Différents critères ont été développés pour évaluer les sous-ensembles candidats. Ils explorent des mesures de distance [5], des mesures statistiques [6] ou plus récemment des mesures probabilistes basées sur l'estimation de l'information mutuelle [7]. Dans cet article, nous proposons d'étudier le potentiel des critères d'information basés sur un calcul du maximum de vraisemblance (MV) pénalisé pour la sélection de variables [8]. Ces critères sont intégrés dans les schémas SFS et SBS. La détermination du MV exige l'estimation de la distribution des variables. En effet, nous adoptons ici le codage arithmétique pour estimer la densité de probabilité et la partition optimale des variables [9].

## 2 Critères d'information

Dans ce travail nous proposons l'adaptation et l'intégration des critères d'information AIC [8], BIC [10] et $\varphi_\beta$ [11] (resp. pour Akaike, Bayes Information Criterion et phi-beta) dans un schéma de sélection de variables. Rappelons que la forme générale d'un critère d'information IC pour un modèle paramétré $Y$ s'écrit sous la forme suivante :

$$IC(Y) = -log(MV) + |Y|\alpha(n) \quad (1)$$

où $|Y|$ est le nombre de paramètres libres qui croît en fonction de la complexité du sous-ensemble et $\alpha(n)$ est une fonction pénalisante qui varie en fonction du critère utilisé (respectivement les $\alpha(n)$ de AIC, BIC et $\varphi_\beta$ sont égaux à 2, $\log n$ et $n^\beta \log\log n$ avec $0 < \beta < 1$ pour assurer de bonnes propriétés asymptotiques). La quantité $|Y|\alpha(n)$ permet de régulariser le comportement du MV car il est connu que ce dernier sur-paramétrise. Ainsi la minimisation du critère permet d'avoir un compromis entre l'adéquation aux données et la complexité des sous-ensembles candidats. En vue de résoudre un problème de sélection de variables à l'aide des ICs, nous nous plaçons dans le cas suivant : soient $X$ un ensemble à $M$ variables et $Y \subseteq X$, sous-ensemble de $X$ à $m$ variables. On dispose de $n$ observations. Chaque observation est assignée à une classe $k$, $k = 1, \cdots, K$. $K$ est le nombre total de classes. La probabilité jointe d'avoir la classe $k$ et l'ensemble $X$ peut être formulée par $P(k, X) = \frac{P(k,Y)\cdot P(X)}{P(Y)}$ dans le cas où la classification devient indépendante du sous-ensemble $X \setminus Y$ conditionnellement à $Y$ [8]. Nous supposons que $P(Y) = 1$ pour $Y = \emptyset$ (sous-ensemble vide). D'après [8], Les ICs sont ainsi définis par

$$IC(Y) = -\sum_{k=1}^{K} \sum_{c \in C_Y} n_{k,c} \log\left\{\frac{n_{k,c}}{n(Y)}\right\} + \left(\left|C_{k,Y}\right| - |C_Y|\right)\alpha(n) \quad (2)$$

où $\left|C_{k,Y}\right|$ et $|C_Y|$ dénotent respectivement le nombre de zones (appelées ensuite « boites ») de l'histogramme multidimensionnel d'occurrence non nul en lien avec le sous-ensemble $Y$, d'une part en prenant compte des différentes classes et d'autre part sans tenir compte de la classification. Nous supposons que $0 \times \log 0 = 0$ et $n(Y) = n$. $n_{k,c}$ est donc le nombre d'observations se trouvant dans la boite $c$ de l'histogramme multidimensionnel. L'obtention de l'histogramme multidimensionnel est réalisée en recherchant au préalable un histogramme optimal pour chaque variable.

## 3 Estimation de la distribution des variables

L'estimation optimale de la densité de probabilité (ddp) est un problème difficile. Cette estimation s'appuie sur un ensemble fini d'observations $v^n = (v_1, \cdots, v_n)$, réalisation de $n$ variables aléatoires indépendantes et identiquement distribuées (iid) $(V_1, \cdots, V_n)$ de ddp $f$ qui est donc à estimer à partir de $v^n$. Nous citons ici les méthodes non paramétriques les plus populaires : les méthodes à noyau et les méthodes par histogramme. Nous nous intéressons particulièrement à l'estimation de la ddp par histogramme en utilisant le codage arithmétique adaptatif [9].

Soient $f$ une ddp inconnue définie sur l'intervalle $I = [a, b] \subset \mathbb{R}$ et $v^n$ une observation. Soit $\mathrm{P} = \left(I_j\right)_{j \in 1,\cdots,m}$ une partition de $I$ à $m \in \mathbb{N}^*$ intervalles. Une estimation de $f$ au sens du MV à partir de $v^n$ peut être obtenue en maximisant l'expression suivante :

$$\hat{f}_{\mathrm{P}} = \sum_{j=1}^{m} \frac{n_j}{n|I_j|} \mathbb{I}_{I_j} \tag{3}$$

où $n_j$ est le nombre de valeurs issues de $v^n$ tombant dans l'intervalle $I_j$, $\left|I_j\right|$ est la longueur de $I_j$ et $\mathbb{I}_{I_j}$ dénote la fonction indicatrice de l'intervalle $I_j$. La difficulté d'une telle procédure d'estimation est de trouver la partition $P$ qui estime au mieux $f$ par $\hat{f}_P$. Pour cela, Coq et al. [9] ont élaboré une technique basé sur les critères d'information et la longueur du codage sans perte des données $v^n$. Cette technique est basée sur deux types de codage : le codage arithmétique et le codage de précision. Nous rappelons ici le critère résultant de ces développements.

$$CRIT(v^n|P) = -\sum_{j=1}^{m} n_j \log \frac{n_j}{n|I_j|} + \frac{(m-1)}{2} \alpha(n) - n \log r \tag{4}$$

Le principe MDL (pour Minimum Description Length) suggère de choisir la partition qui minimise ce critère pour approcher $f$ par $\hat{f}_{\hat{P}}$ de la manière suivante :

$$\hat{P} = argmin\{CRIT(v^n, P), P \in P_{max}\} \tag{5}$$

où $P_{max}$ est la partition maximale de $I$ à $R$ intervalles tous de longueur $r$ définie à *priori*.

L'obtention d'une telle partition et d'une telle densité pour chaque variable à l'aide de la programmation dynamique afin de ne pas tester toutes les partitions possibles constitue le point de départ pour calculer l'équation (2).

## 4 Expérimentations

Notre approche de sélection de variables est basée sur les critères d'information définis par l'équation $(2)$. Elle est évaluée et comparée d'abord sur des données simulées qui suivent une loi normale $\mathcal{N}(\mu, \sigma)$ où le nombre de classes $K$ est égal à 2, le nombre de variables $M = 3$ et $n = 500$ par classe

(figure 1), ensuite sur le jeu de données WINE[1] dont les caractéristiques sont $K = 3$, $n = 178$, et $M = 13$). Pour ce jeu de données, nous avons utilisé la validation croisée (CV pour cross validation) pour construire les ensembles d'apprentissage et de test $(10 - CV)$. Cette opération est répétée $N$-fois pour obtenir des résultats statistiquement corrects $(N = 100)$.

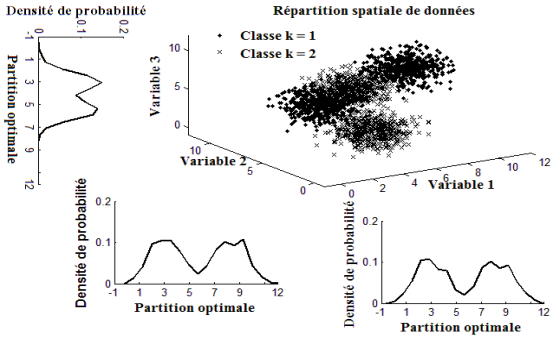


**Figure 1 :** Répartitions spatiales des données et ddp. Les variables suivent une loi normale $\mathcal{N}(\mu, \sigma = 1)$ sur chaque axe. Le tableau 1 détaille les valeurs de la moyenne $\mu$ pour chaque variable et chaque classe $(K = 2)$.

**Tableau 1 :** Valeurs de moyenne $\mu$ prisent pour chaque variable en fonction de chaque classe.

| Numéro classe | Variable 1 | Variable 2 | Variable 3 |
|---|---|---|---|
| $k = 1$ | $(2,10)$ | $(4,8)$ | $(6,6)$ |
| $k = 2$ | $(4,8)$ | $(2,10)$ | $(3,3)$ |

A partir des ddp estimées en utilisant l'équation (4), nous avons calculé les ICs de l'équation (2) en utilisant les schémas de parcours SFS et SBS. Le but est de quantifier la qualité des variables afin de sélectionner celles qui décrivent au mieux les données. Le tableau 2 montre les sous-ensembles sélectionnés à chaque itération par le critère d'information BIC dans le cas de la figure 1. Ces résultats sont comparés au critère de Wilk's tel qu'il est défini dans [12]. En analysant uniquement les ddp, on observe que le sous-ensemble formé uniquement par la variable 3 décrit bien les données. Un sous-ensemble contenant cette variable ne peut que mieux représenter ces données. Cependant, le sous-ensemble $\{1,2\}$ sépare presque parfaitement les données. Le tableau 2 montre aussi la différence du sous-ensemble obtenu à l'itération 2 par le schéma SBS entre les critères BIC ($Y_{SBS} = \{1,2\}$) et Wilk's ($Y_{SBS} = \{2,3\}$). Cependant, ces critères, pour le schéma SFS, renvoi le même sous ensemble ($Y_{SFS} = \{2,3\}$). D'où l'intérêt du schéma SBS en utilisant le critère BIC.

**Tableau 2 :** Résultats de sous-ensembles sélectionnés.

| | SFS | | SBS | |
|---|---|---|---|---|
| | BIC | Wilk's | BIC | Wilk's |
| $Itération\ 1$ | $\{3\}$ | $\{3\}$ | $\{1,2,3\}$ | $\{1,2,3\}$ |
| $Itération\ 2$ | $\{2,3\}$ | $\{2,3\}$ | $\{1,2\}$ | $\{2,3\}$ |
| $Itération\ 3$ | $\{1,2,3\}$ | $\{1,2,3\}$ | $\{2\}$ | $\{3\}$ |

[1] Base de données UCI Repository : http://archive.ics.uci.edu/ml/datasets.html

La figure 2 ci-dessous présente une comparaison quantitative entre les différents sous-ensembles obtenus par les trois critères d'information et le critère de Wilk's. Ces résultats montrent les pourcentages moyens de taux de bonne classification obtenus par les algorithmes de classification. Quatre algorithmes de classification ont été appliqués pour mesurer la qualité des résultats et les performances des schémas SFS et SBS à savoir $k$-NN ($k = 1$, $k$-Neareast Neighbor), NB (Naive Bayes), LDA (Linear Discriminant Analysis) et SVM (Support Vector Machine).

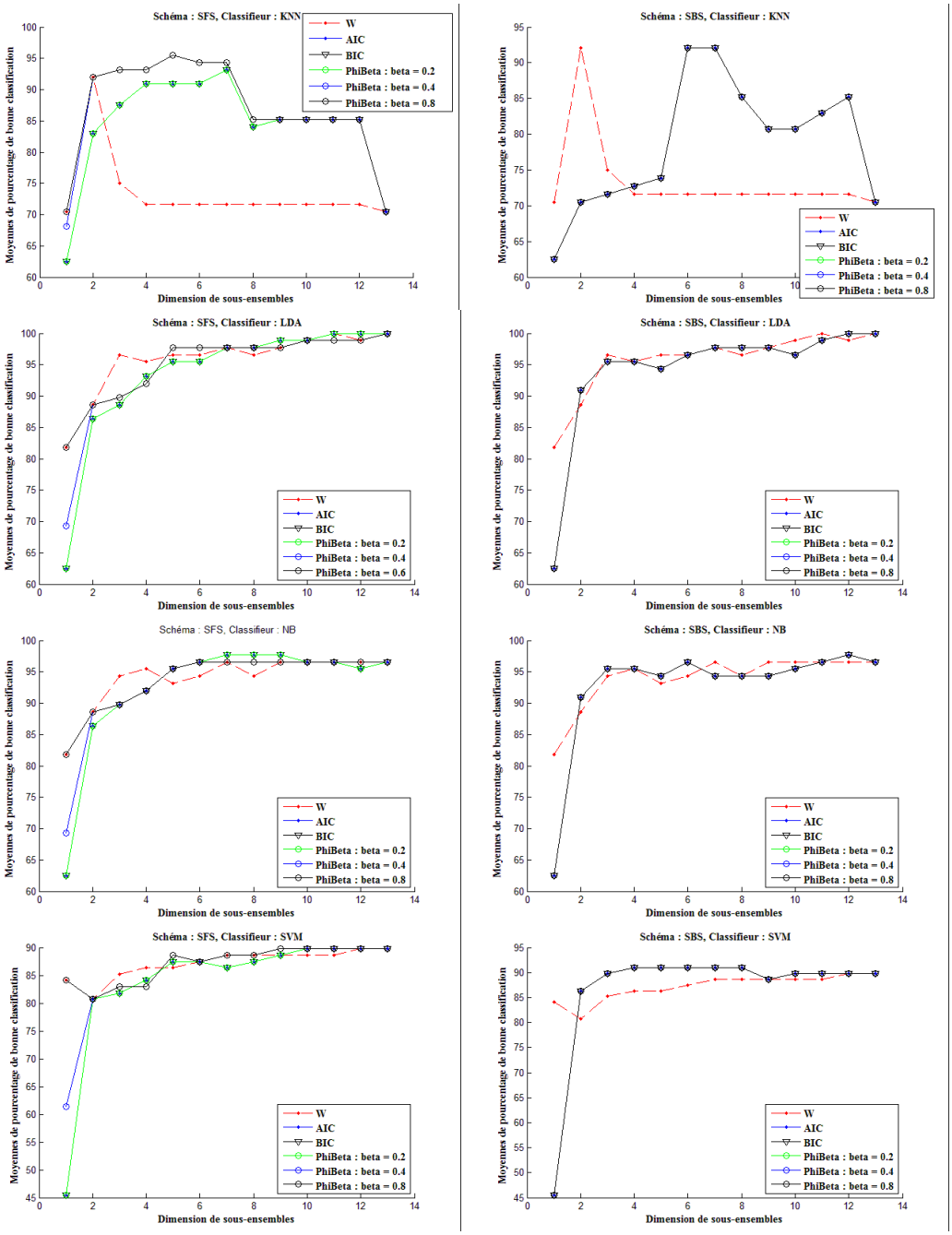


**Figure 2 :** Comparaison de taux de bonne classification entre les critères d'information et celui de Wilk's pour le jeu de données WINE. Par lignes classifieurs $k$-NN, LDA, NB et SVM. Par colonnes schémas SFS et SBS.

La figure 2 montre la supériorité des critères d'information par rapport au critère de Wilk's. Les courbes issues des critères d'information se confondent dans la plupart de cas. Cela signifie que ces critères ont sélectionné les mêmes sous-ensembles pour les différentes itérations. Pour le critère $\varphi_\beta$ nous avons choisi $\beta = 0.2,\ 0.4$ et $0.8$. Pour ces valeurs de $\beta$, nous constatons que le taux de bonne classification est le plus élevés. L'utilisation de différents algorithmes de classification affecte peu notre approche. Elle converge mieux vers la solution optimale obtenue par une recherche exhaustive.

## 5 Conclusion

Nous avons présenté une approche de sélection de variables basée sur les critères d'information. Des estimateurs continus par morceaux utilisant le codage arithmétique adaptatif sont employés pour approcher la ddp et définir ainsi la partition optimale pour chaque variable. Les tests réalisés sur des données simulées et références et les comparaisons avec le critère de Wilk's ont montré la bonne tenue de notre approche pour différents algorithmes de classification.

## Remerciements

Les auteurs remercient les institutions européennes à travers le programme INTERREG IV (projet CESAR volet II) dont l'Université des Antilles et de la Guyane est partenaire et la région Poitou-Charentes d'avoir financé ce projet.

## Références

[1] A.K. Jain, R.P.W. Duin, and J. Mao. Statistical pattern recognition: A review. IEEETrans. PAMI, 22(1) :4-37, 2000.

[2] S. Mitraetal. Datamining in soft computing frame work: A survey. IEEETrans.NN, 13(1) :3-14,2002.

[3] H. Liuand and L.Yu. Toward integrating FS algorithms for classification and clustering. IEEETrans. on KDE, 17(4) :491–502, 2005.

[4] Y. Sun, S. Todorovic, and S. Goodison. Local-Learning-Based Feature Selection for High-Dimensional Data Analysis. TPAMI, 32,(9) :1610-1626, 2010.

[5] J. Liang, Su Yang and A-C. Winstanley. Invariant optimal feature selection: A distance discriminant and feature ranking based solution. Pattern Recognition, 41(5) :1429-1439, 2008.

[6] G. Qu, S. Hariri and M. Yousif. A new dependency and correlation analysis for features. IEEE Transactions on Knowledge and Data Engineering, 17(9) :1199-1207, 2005.

[7] H. Liu, J. Sun, L. Liuand and H. Zhang. Feature selection with dynamic mutual information. Pattern Recognition, 42(7) :1330-1339, 2009.

[8] Y. Sakamoto and H. Akaike. Analysis of cross classified data by AIC. Ann. Inst. Statist. Math, 30(B) :185-197, 1978.

[9] G. Coq, O. Alata, Y. Pousset, X. Li and C. Olivier. Law recognition via histogram-based estimation. IEEE ICASSP, Taïpei (Taïwan), 3425-3428, april 2009.

[10] T. Mary-Huard and E. Lebarbier. Une introduction au critère BIC : fondements théoriques et applications. Journal de la société française de Statistiques, 147(1) :39-57, 2006.

[11] D.A. El-Matouat and M. Hallin. Order selection, stochastic complexity and Kullback-Leibler information. 115 :291-299, 1996.

[12] A. Porebski, N. Vandenbroucke and L. Macaire. Comparison of feature selection schemes for color texture classification. IPTA, Paris (France), july 2010.