Arbre BIC optimal et taux d'erreur

Gilbert Ritschard

Département d'économétrie, Université de Genève
gilbert.ritschard@themes.unige.ch

Résumé. Nous reconsidérons dans cet article le critère BIC pour arbres d'induction proposé dans Ritschard et Zighed (2003, 2004) et discutons deux aspects liés à sa portée. Le premier concerne les possibilités de le calculer. Nous montrons comment il s'obtient à partir des statistiques du rapport vraisemblance utilisées pour tester l'indépendance ligne-colonne de tables de contingence. Le second point porte sur son intérêt dans une optique de classification. Nous illustrons sur l'exemple du Titanic la relation entre le BIC et le taux d'erreur en généralisation lorsqu'on regarde leur évolution selon la complexité de l'arbre. Nous esquissons un plan d'expérimentation en vue de vérifier la conjecture selon laquelle le BIC minimum assurerait en moyenne le meilleur taux d'erreur en généralisation.

1 Introduction

La qualité des arbres de classification, comme pour d'autres classifieurs, est le plus souvent établie sur la base du taux d'erreur de classement en généralisation. Si l'on examine l'évolution de ce taux en fonction de la complexité du classifieur, il est connu qu'il passe par un minimum au delà duquel on parle de sur-apprentissage (*overfitting*). Intuitivement, l'explication de ce phénomène tient au fait qu'au delà d'un certain seuil, plus on augmente la complexité, plus l'arbre devient dépendant de l'échantillon d'apprentissage utilisé, au sens où il devient de plus en plus probable que de petites perturbations de l'échantillon entraîneront des modifications des règles de classification. Lorsqu'il s'agit d'utiliser l'arbre pour la classification, il semble dès lors naturel de retenir celui qui minimise le taux en généralisation.

Mais comment s'assurer a priori que l'arbre induit sera celui qui minimisera le taux en généralisation ? On pourrait songer à partager les données disponibles pour l'induction en un échantillon d'apprentissage et un échantillon test et à exploiter le taux d'erreur sur les données test comme critère de construction de l'arbre. Ceci reviendrait cependant simplement à transformer les données test en données d'apprentissage et ne peut donc être une solution. Il s'agit de disposer d'un critère qui, tout en se calculant sur l'échantillon d'apprentissage, nous assure que le taux d'erreur sera en moyenne minimum pour tout ensemble de données supplémentaires. A défaut de pouvoir mesurer a priori le taux d'erreur en généralisation, on s'intéresse à la complexité qu'il s'agit de minimiser et l'on tentera de retenir le meilleur compromis entre qualité d'information sur données d'apprentissage et complexité.

Le critère BIC (Bayesian Information Criteria) pour arbre que nous avons introduit dans Ritschard et Zighed (2003, 2004) pour comparer la qualité de la description

des données fournies par différents arbres nous semble pouvoir être une solution de ce point de vue puisqu'il combine un critère d'ajustement (la déviance) avec une pénalisation pour la complexité (le nombre de paramètres). D'autres critères, dont la description minimale de données (Rissanen, 1983) et le message de longueur minimal, MML, (Wallace et Freeman, 1987) qui combinent également une qualité d'information et une pénalisation pour la complexité pourraient également s'avérer intéressants de ce point de vue. Le critère BIC considéré ici résulte d'une logique bayésienne (Raftery, 1995) tout comme le critère que Wehenkel (1993) utilise pour l'élagage.

Avant d'examiner le lien du BIC avec le taux d'erreur en généralisation, nous rappelons à la section 2 sa définition et en particulier celle de la déviance sur la laquelle il se fonde. Nous montrons que la déviance se déduit directement de la valeur de la statistique du rapport de vraisemblance de deux tests d'indépendance, ce qui permet en particulier à tout un chacun de calculer le BIC d'un arbre en utilisant n'importe quel logiciel statistique classique, SPSS par exemple, qui donne ces statistiques. Nous illustrons ensuite à la section 3 le lien entre le critère BIC et le taux d'erreur et discutons brièvement d'un protocole d'expérimentation en vue de vérifier la conjecture selon laquelle la minimisation du BIC assurerait la minimisation du taux moyen d'erreur en généralisation.

2 Le critère BIC pour arbre d'induction

Pour illustrer notre discussion nous utilisons les données du Titanic où il s'agit de prédire pour chaque passager s'il survit ou pas selon trois attributs, soit le sexe (F,M), l'âge (A=adulte, C=enfant) et la classe (c1, c2, c3 et c4=équipage). La figure 1 donne l'arbre induit avec la méthode Exhaustive CHAID de Answer Tree 3.1 (SPSS, 2001) en utilisant le khi-deux du rapport de vraisemblance, un seuil de signification de 5% et les contraintes minimales sur la taille des nœuds.

Avant de définir le critère BIC, nous devons expliquer la déviance que nous notons $D(m)$ pour un arbre m. Considérons pour cela la table de contingence cible dont les ℓ lignes sont définies par la variable à prédire ("survit ou pas" dans notre cas) et dont les c colonnes correspondent à l'ensemble des profils différents que l'on peut définir avec les attributs prédictifs. Dans notre cas on a 14 profils différents, soit $2 \times 2 \times 4$ moins 2 puisqu'il n'y a pas d'enfant, ni fille ni garçon parmi l'équipage. La déviance mesure la divergence entre la table cible et sa prédiction à partir de l'arbre induit. Formellement, la déviance se calcule comme suit

$$D(m) \quad = \quad -2 \sum_{i=1}^{\ell} \sum_{j=1}^{c} n_{ij} \ln \left(\frac{\hat{n}_{ij}}{n_{ij}} \right) \tag{1}$$

en considérant les termes $n_{ij} \ln(\hat{n}_{ij}/n_{ij})$ comme nuls lorsque $n_{ij} = 0$.

Les prédictions $\hat{n}_{ij}$ s'obtiennent en ventilant le total des cas avec profil j selon la distribution observée dans la feuille de l'arbre induit qui comprend le profil j. Le tableau 1 donne par exemple (sous forme transposée pour des raisons de présentation) la table cible et la table prédite avec l'arbre induit de la figure 1. Avec la formule ci-dessus on établit que la déviance vaut ici 6.93.

	table cible (n_{ij})		table prédite ($\hat{n}_{ij}$)		
	yes	no	yes	no	total
MAc1	45	88	45	88	133
MAc2	10	114	10	114	124
MAc3	59	289	67.6175	280.3825	348
MAc4	132	503	123.3825	511.6175	635
MCc1	4	0	4	0	4
MCc2	8	0	8	0	8
MCc3	10	23	10	23	33
FAc1	112	4	112.0342	3.9658	116
FAc2	66	11	67.375	9.625	77
FAc3	62	71	59.5	73.5	133
FAc4	15	2	15	2	17
FCc1	1	0	0.9658	0.0342	1
FCc2	11	0	9.625	1.375	11
FCc3	6	13	8.5	10.5	19
total	541	1118	541	1118	1659

TAB. 1 – Table cible et effectifs prédits

Le critère BIC pour un arbre induit m qui donne lieu à $q \leq c$ feuilles pour une variable à prédire avec ℓ classes est alors défini, à une constante additive près, par

$$\text{BIC}(m) \quad = \quad D(m) + p\ln(n) \tag{2}$$

où n est la taille de l'échantillon d'apprentissage, $p = (\ell-1)q+c$ le nombre de paramètres de l'arbre et $D(m)$ la déviance.

En présence d'un grand nombre d'attributs, le nombre de profils différents possibles peut évidemment rapidement devenir trop grand pour envisager un calcul manuel de

	table T_m		
	yes	no	total
MAc1	45	88	133
MAc2	10	114	124
MAc3,c4	191	792	983
MCc1	4	0	4
MCc2	8	0	8
MCc3	10	23	33
FA,Cc1	113	4	117
FA,Cc2	77	11	88
FA,Cc3	68	84	152
FAc4	15	2	17
total	541	1118	1659

TAB. 2 – Table croisant la variable à prédire avec les feuilles

Arbre BIC optimal et taux d'erreur

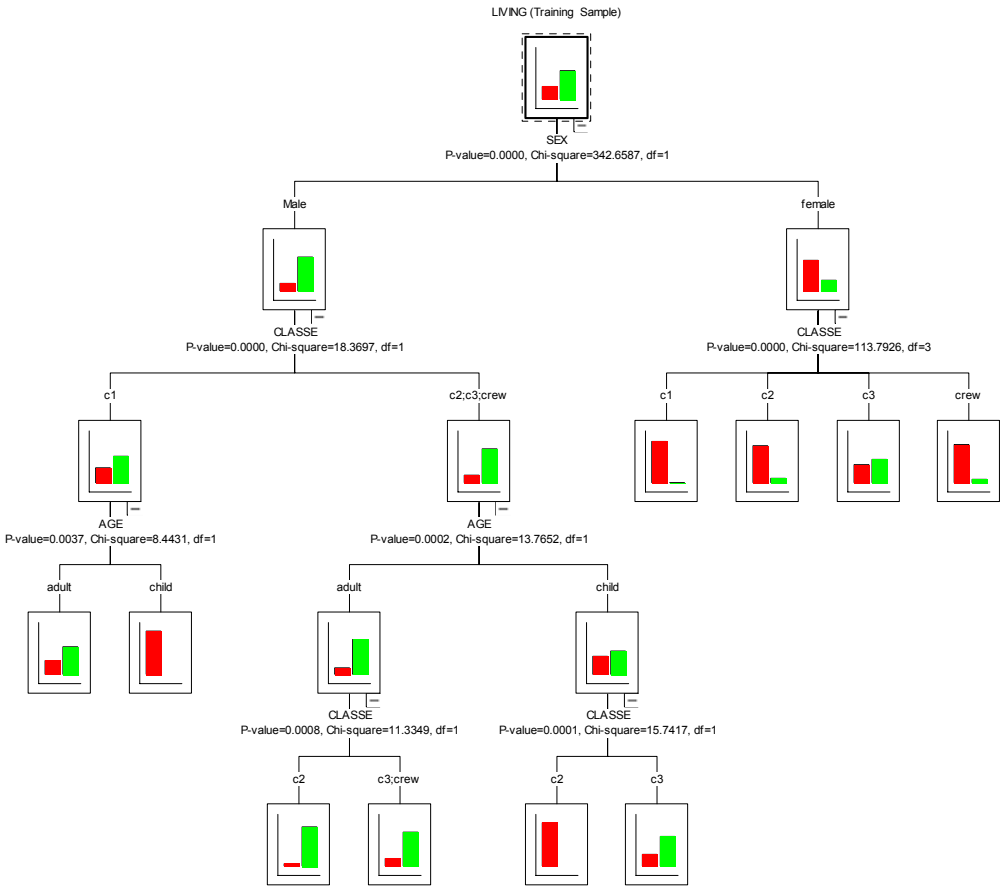

FIG. 1 – Arbre induit, échantillon d'apprentissage, $n = 1659$

la déviance. On peut dans ce cas exploiter les logiciels qui calculent la statistique du rapport de vraisemblance pour le test d'indépendance sur table de contingence. Notons T la table cible et T_m la table de contingence qui croise la variable à prédire avec les feuilles de l'arbre induit. Pour notre exemple, on trouve T au tableau 1 tandis que T_m est donné au tableau 2. En raison des propriétés d'additivité de la déviance, on a $D(m_0) = D(m) + D(m_0|m)$ où m_0 représente le nœud initial qui correspond à l'indépendance. La divergence $D(m_0)$ entre ce nœud initial et la table cible est précisément la statistique du rapport de vraisemblance pour le test d'indépendance sur le tableau T. De même la divergence $D(m_0|m)$ entre m_0 et m est le khi-deux du rapport de vraisemblance pour le test d'indépendance sur la table T_m.

La statistique du rapport de vraisemblance vaut respectivement 531.04 pour T et 524.11 pour T_m. On trouve donc $D(m) = 531.04 - 524.11 = 6.93$ ce qui correspond à la valeur trouvée précédemment. On en déduit la valeur du BIC, soit $\text{BIC}(m) = 6.93 + 24 \ln(1659) = 184.87$.

Notons que le BIC est défini à une constante additive près. Comme à chaque fois que le nombre de paramètres augmente d'une unité le nombre d de degrés de liberté associé à la déviance diminue d'une unité, on peut également considérer la définition $\text{BIC}(m) = D(m) - d\ln(n)$. C'est la définition que nous avons utilisée au tableau 3 où nous avons cependant encore ajouté 100 pour éviter les valeurs négatives.

3 BIC et le taux d'erreur en généralisation

On se propose à présent de discuter le lien entre le critère BIC et le taux d'erreur en généralisation. Notre conjecture est que l'arbre BIC optimal devrait plus ou moins assurer le plus petit taux d'erreur en généralisation.

Nous avons calculé la valeur du critère BIC et le taux d'erreur pour différents arbres obtenus en élaguant successivement l'arbre saturé de la figure 3. L'échantillon d'apprentissage comprend 1659 cas et l'échantillon test 542. On notera que sur cet exemple très simple, le taux d'erreur en généralisation, bien que supérieur au taux d'erreur en substitution, ne remonte pas pour les arbres les plus complexes. Le tableau 3 et la figure 2 font apparaître que le BIC correspond au modèle le plus simple pour lequel on a le taux d'erreur minimal. Ceci semble plutôt conforter notre conjecture.

Nous n'avons ici bien évidemment considéré qu'un seul ensemble test qui ne saurait suffire à démontrer notre conjecture. Notre intention est de procéder à une expérimentation plus complète. Le protocole envisagé est de postuler successivement plusieurs structures et de générer pour chacune d'entre elle un échantillon d'apprentissage et un ensemble de disons 100 échantillons tests. Comme dans l'exemple ci-dessus, nous considérerons plusieurs arbres de complexité variable pour lesquels nous calculerons le critère BIC. Chaque arbre sera ensuite appliqué sur chacun des échantillons test, et nous compareront la moyenne des taux d'erreur obtenus avec le critère BIC. Plus précisément nous comparerons l'évolution avec la complexité de ces deux indicateurs.

						taux d'erreur	
regroup.	model	p	-2LL	d	BIC	test	apprent.
	saturated	14	0	0	100	0.221	0.206
A,C | F,c1	m1	13	0.07	1	92.66	0.221	0.206
A,C | F,c3	m2	12	1.63	2	86.81	0.221	0.206
c3,c4 | M,A	m3	11	3.78	3	81.54	0.221	0.206
A,C | F,c2	m4	10	6.93	4	77.28	0.221	0.206
A,C | M,c1	m5	9	15.37	5	78.30	0.223	0.208
c2,c3c4 | M,A	m6	8	26.71	6	82.23	0.223	0.208
c2,c3 | M,C	m7	7	42.45	7	90.55	0.229	0.213
A,C | M,c2c3c4	m8	6	56.22	8	96.90	0.229	0.213
c1,c2c3c4 | M	m9	5	74.59	9	107.86	0.229	0.213
c1,c2,c3,c4 | F	m10	2	188.38	12	199.41	0.229	0.222
tout	indep	1	531.04	13	534.66	0.314	0.326

Tab. 3 – Qualité des modèles successifs

Arbre BIC optimal et taux d'erreur

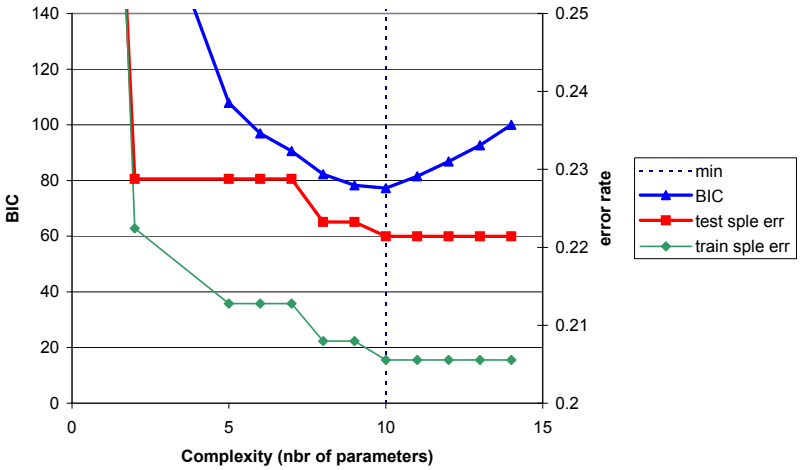

FIG. 2 – BIC sur échantillon d'apprentissage et taux d'erreur

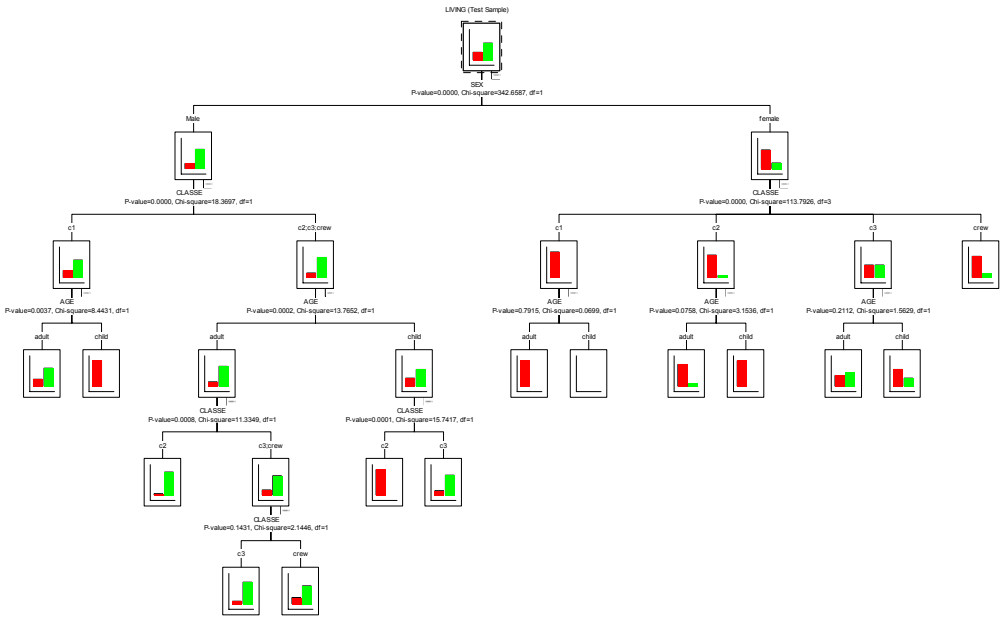

FIG. 3 – Arbre saturé, échantillon test, $n = 542$

4 Conclusion

Le critère BIC pour arbres a été introduit dans Ritschard et Zighed (2003, 2004) comme critère pour déterminer l'arbre le plus adéquat du point de vue de la description de données. Nous pensons cependant que ce critère peut également s'avérer utile dans

une optique de classification. L'illustration considérée semble confirmer notre conjecture selon laquelle le critère BIC devrait permettre de déterminer l'arbre le mieux adapté à une utilisation prédictive en dehors de l'échantillon d'apprentissage. Une expérimentation complète s'avère cependant nécessaire pour donner une assise empirique mieux fondée à cette conjecture.

Références

Raftery, A. E. (1995). Bayesian model selection in social research. In P. Marsden (Ed.), *Sociological Methodology*, pp. 111–163. Washington, DC : The American Sociological Association.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *The Annals of Statistics 11*(2), 416–431.

Ritschard, G. et D. A. Zighed (2003). Goodness-of-fit measures for induction trees. In N. Zhong, Z. Ras, S. Tsumo, et E. Suzuki (Eds.), *Foundations of Intelligent Systems, ISMIS03*, Volume LNAI 2871, pp. 57–64. Berlin : Springer.

Ritschard, G. et D. A. Zighed (2004). Qualité d'ajustement d'arbres d'induction. *Revue des nouvelles technologies de l'information E-1*, 45–67.

SPSS (Ed.) (2001). *Answer Tree 3.0 User's Guide*. Chicago : SPSS Inc.

Wallace, C. S. et P. R. Freeman (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society, Series B (Methodological) 49*(3), 240–265.

Wehenkel, L. (1993). Decision tree pruning using an additive information quality measure. In B. Bouchon-Meunier, L. Valverde, et R. Yager (Eds.), *Uncertainty in Intelligent Systems*, pp. 397–411. Amsterdam : Elsevier - North Holland.

Summary

We discuss two aspects related to the scope of the BIC index for induction trees proposed in Ritschard et Zighed (2003, 2004). The first point is about how to compute it. We show that the BIC can easily be derived from the Likelihood Ratio Chi-square statistics used for testing the row-column independence of contingency tables. The second aspect is related to its interest for classification purposes. We illustrate, by means of the Titanic example, the expected link between the BIC and the generalization error rate in terms of their evolution with respect to the tree complexity. Finally, we sketch an experiment design for checking empirically the conjecture that the minimal BIC ensures on average the best generalization error rate.