france telecom

Recherche & Développement

NT/FT/R&D/8611

Juillet 2004

TECH
Technologies

MODL : une méthode quasi-optimale de groupage des valeurs d'un attribut symbolique

Marc BOULLE (TECH/SUSI)

Note Technique

diffusion libre

& france telecom

Recherche & Développement

Note Technique
NT/FTR&D/8611

Juillet 2004

MODL: une méthode quasi-optimale de groupage des valeurs d'un attribut symbolique

Marc Boullé (TECH/SUSI)

Vu, pour accord le
Directeur de TECH

P. A. Badoz

Vu, le responsable du
laboratoire SUSI

V.Beaudouin

Date : 19 juillet 2004

Résumé : Dans le domaine de l'apprentissage supervisé, les méthodes de groupage des valeurs d'un attribut symbolique partitionnent l'ensemble des valeurs de l'attribut en un nombre fini de groupes, en recherchant un compromis entre valeur informationnelle et valeur prédictive de la partition formée. Dans ce document, nous introduisons une formalisation du problème du groupage supervisé. Dans ce cadre nous proposons un critère d'évaluation d'un groupage, dont l'optimisation garantit l'optimalité au sens de Bayes du groupage obtenu. Nous présentons également des algorithmes permettant une optimisation poussée des groupages. Des expérimentations comparatives intensives montrent que la méthode MODL* est très performante tant en termes de qualité prédictive, de robustesse que de compacité des groupages.
*MODL : Minimum Optimized Description Length

Mots clés : **analyse intelligente donnée ; apprentissage automatique ; groupage ; méthode Bayes**
Domaine : Traitement de l'information et des connaissances

France Télécom
Recherche et Développement
2, avenue Pierre Marzin – 22307 Lannion Cedex
Téléphone : 02 96 05 11 11
Téléphone international : +33 2 96 05 11 11
SA au capital de 9 609 262 400 € - 380 129 866 RCS Paris

MODL: une méthode quasi-optimale de groupage des valeurs d'un attribut symbolique

MARC BOULLE
France Telecom R&D
2, Avenue Pierre Marzin
22300 Lannion – France
marc.boulle@francetelecom.com

Résumé. Dans le domaine de l'apprentissage supervisé, les méthodes de groupage des valeurs d'un attribut symbolique partitionnent l'ensemble des valeurs de l'attribut en un nombre fini de groupes, en recherchant un compromis entre valeur informationnelle et valeur prédictive de la partition formée. Dans ce papier, nous introduisons une formalisation du problème du groupage supervisé. Dans ce cadre nous proposons un critère d'évaluation d'un groupage, dont l'optimisation garantit l'optimalité au sens de Bayes du groupage obtenu. Dans le cas à deux classes cibles, nous montrons que les algorithmes de discrétisation MODL sont applicables, ce qui permet de trouver le groupage optimal en $O(I^3)$ (où I est le nombre de valeurs descriptives à grouper), ou un groupage quasi-optimal au moyen d'une heuristique en $O(I.\log(I))$. Dans le cas général avec un nombre de classes cibles supérieur à deux, nous présentons de nouveaux algorithmes de groupage en $O(I.\log(I))$ incorporant des étapes de pré-optimisation et post-optimisation. Des expérimentations intensives ont été menées sur de nombreux jeux de données de nature différente afin de comparer la méthode de groupage MODL avec d'autre méthodes de référence, en prenant en compte la performance prédictive, la robustesse et la taille des groupages. Une analyse multicritères des résultats démontre les bonnes performances de la méthode MODL, tout en apportant un éclairage intéressant sur la problématique générale du groupage.

Mots clés : Data Mining, Machine Learning, Value Grouping, Data Analysis

TABLE DES MATIERES

1 Introduction

Les méthodes d'apprentissage supervisé sont au cœur de l'étape de modélisation du Data Mining. Elles consistent à prédire les valeurs d'un attribut cible (également appelées classes) à partir d'un ensemble d'attributs descriptifs, de nature numérique ou symbolique. Le problème du groupage des modalités (ou valeurs descriptives) d'un attribut symbolique consiste à partitionner l'ensemble des valeurs de l'attribut en un nombre fini de groupes identifiés chacun par un code. La plupart des modèles prédictifs à base d'arbres de décision utilisent une méthode de groupage pour traiter les attributs symboliques, de façon à lutter contre la fragmentation des données. Les méthodes à base de réseaux de neurones n'utilisant que des données numériques ont souvent recours à un codage disjonctif complet des attributs symboliques. Dans le cas où les modalités sont trop nombreuses, il est nécessaire de procéder au préalable à des regroupements de modalités. Ce problème se rencontre également dans le cas des réseaux bayésiens ou de la régression logistique. De façon générale, le groupage est une technique intéressante de préparation des données pour le Data Mining, qui permet d'identifier les groupes de modalités homogènes vis à vis de l'attribut à prédire.

Les méthodes de groupage peuvent se catégoriser en fonction de la stratégie de recherche du meilleur groupage et du type de critère d'évaluation à optimiser. Plusieurs stratégies de groupage ont été explorées dans la bibliographie. L'algorithme le plus simple est la binarisation où une modalité est isolée contre toutes les autres. Une stratégie plus élaborée consiste à rechercher le regroupement optimal des modalités en deux groupes (optimal au sens du critère optimisé). L'algorithme Sequential Forward Selection inspiré de (Cestnik, Kononenko & Bratko 1987) et évalué par (Berckman 1995) est un algorithme glouton qui recherche la meilleure bipartition des modalités en déplaçant les modalités une à une d'un premier groupe initialement complet vers un second groupe initialement vide. Dans le cas d'un problème à deux classes (i.e. deux classes cibles), (Breiman, Friedman, Olshen et Stone 1984) ont présenté un algorithme optimal de regroupement en deux parties des modalités pour certaines familles de critère. Cet algorithme est basé sur un tri préalable des modalités par proportion croissante de la première classe cible, puis sur le choix d'une coupure entre deux modalités adjacentes dans cette liste triée de modalités. La complexité de cet algorithme est en $I.\log(I)$ où I est le nombre de valeurs descriptives initiales. En se basant sur les idées de (Lechevallier 1990; Fulton, Kasif & Salzberg 1995), il est possible d'étendre ce résultat à une partition optimale en K groupes dans le cas de deux classes cibles, en utilisant un algorithme de programmation dynamique de complexité quadratique par rapport au nombre I de valeurs descriptives initiales. Dans le cas général, il n'existe pas d'algorithme de recherche de groupage optimal autre que la recherche exhaustive, qui n'est pas envisageable. (Chou 91) a néanmoins mis en évidence des conditions d'optimalité permettant de réduire l'espace de recherche, et proposé un algorithme de type K-means permettant de trouver une K-partition des modalités localement optimale. La complexité algorithmique est en $K*I$ multiplié par le nombre d'itérations (en général faible), mais l'optimalité globale n'est pas assurée et le nombre K de groupes est un paramètre utilisateur. En pratique, la stratégie de groupage des valeurs descriptives repose souvent sur l'utilisation d'un algorithme glouton itératif (Kass 1980; Quinlan 1993). Cet algorithme est similaire à une classification hiérarchique ascendante des modalités et regroupe itérativement les modalités pour optimiser un critère de qualité du groupage, en s'arrêtant quand le maximum est atteint (ce qui détermine automatiquement le nombre K de groupes). (Ritschard, Zighed et Nicoloyannis 2001) ont comparé cet algorithme glouton avec une recherche exhaustive optimale pour le critère de Tschuprow appliqué à des jeux d'essai artificiels de petite taille, dans le cas de regroupement de lignes et de colonnes d'une table de contingence. Ils ont montré que l'algorithme glouton trouvait des solutions très proches de la solution optimale.

Les critères utilisés pour évaluer la qualité d'un groupage sont très nombreux: il s'agit en fait des critères utilisés pour évaluer une table de contingence. L'algorithme ID3 (Quinlan 1986) utilise le gain informationnel basé sur l'entropie de Shannon pour comparer l'importance prédictive des attributs, sans procéder à des regroupements de modalités. Ce critère favorisant les attributs ayant de nombreuses modalités, (Quinlan 1993) a apporté un correctif heuristique au gain informationnel, le "gain ratio", en divisant le gain informationnel par la quantité d'information contenue dans l'attribut symbolique descriptif. L'algorithme Cart (Breiman 1984) recherche pour chaque attribut une bipartition des modalités en utilisant l'indice de Gini. Dans le cas à deux classes, l'algorithme permet de trouver la bipartition optimale. Dans le cas général, l'algorithme utilise une méthode de recherche exhaustive en évaluant toutes les bipartitions pour l'indice de Gini à L classes. Il propose également un critère alternatif appelé critère Twoing, pour lequel il envisage toutes les bipartitions des classes cibles, et pour chacune recherche la meilleure bipartition des valeurs descriptives en se ramenant à l'indice de Gini à deux classes. La complexité de cette recherche (optimale) étant exponentielle en fonction du nombre de modalités, cette méthode n'est envisageable que dans le cas où il y a peu de valeurs descriptives ou de classes. L'algorithme CHAID (Kass 1980) utilise une méthode de groupage des modalités apparentée à ChiMerge (Kerber 1991). Il s'agit de rechercher la meilleure fusion de modalités en minimisant le critère du Khi2 local aux deux valeurs descriptives candidates, de façon à favoriser le regroupement de modalités ayant un comportement statistique similaire. L'utilisation du critère du Khi2 a également été envisagée pour l'évaluation globale du tableau de contingence et non de façon locale à deux lignes de ce tableau de contingence comme dans CHAID. Dans le cas de l'évaluation globale, les coefficients de Cramer ou de Tschuprow permettant de normaliser la valeur du Khi2 ont été utilisés comme critère de groupage à optimiser. La méthode Khiops (Boullé 2003) utilise le niveau de confiance associé au test du Khi2 pour évaluer les tables de contingence et propose un contrôle statistique de l'algorithme de groupage permettant de fiabiliser la qualité prédictive des groupages. Pour une description approfondie des méthodes à base d'arbre ou de graphe d'induction et de la façon dont elles traitent les attributs symboliques,

on peut se référer à (Zighed et Rakotomalala 2000).

L'enjeu du regroupement des modalités est de trouver une partition réalisant un compromis entre qualité informationnelle (groupes homogènes vis-à-vis de l'attribut à prédire) et qualité statistique (effectifs suffisants pour assurer une généralisation efficace). Ainsi, le cas extrême d'un attribut ayant autant de modalités que d'instances est inutilisable : tout regroupement des modalités correspond à un apprentissage « par cœur » inutilisable en généralisation. Dans l'autre cas extrême d'un attribut réduit à un seul groupe, la capacité en généralisation est optimale, mais l'attribut ne possède aucune information permettant de séparer les classes à prédire. Il s'agit alors de trouver un critère mathématique permettant d'évaluer et de comparer des partitions de taille différente, puis un algorithme conduisant à la meilleure partition.

Nous présentons dans ce papier une nouvelle méthode de groupage appelée MODL, reprenant les principes utilisés par la méthode de discrétisation MODL (Boullé 2004). La principale innovation réside dans le choix du critère d'évaluation d'une partition résultant d'une approche bayésienne du groupage. Le critère d'évaluation utilisé par la méthode MODL a été élaboré d'une part en fonction de ses qualités d'évaluation fine des groupages, y compris sur des échantillons de petite taille ou avec des modalités rares, d'autre part en raison de sa décomposabilité sur l'ensemble des groupes, qui permet d'utiliser des algorithmes d'optimisation performants. Après une formulation explicite du problème de groupage, nous démontrons que le critère proposé permet de trouver le groupage optimal au sens de Bayes, c'est à dire le groupage le plus probable expliquant les données. Nous présentons une heuristique gloutonne ascendante classique en $O(I^2\log(I))$ ou I est le nombre de valeurs à grouper. Nous proposons des améliorations pour cet algorithme sous la forme de pré-optimisations et post-optimisations, de façon à réduire la complexité à $O(I.\log(I))$ tout en améliorant la qualité des groupages. Nous introduisons également une prise en compte des très grands nombres de valeurs à grouper par une fusion préalable des valeurs rares dans un groupe "poubelle". L'effectif maximal des valeurs rares à fusionner dans le groupe poubelle est pris en compte de façon optimale dans le critère d'évaluation des groupages, puis les algorithmes sont adaptés en conséquence avec une dégradation minimale des performances en temps de calcul.

Afin d'évaluer la méthode MODL et de la comparer à plusieurs autres méthodes de groupage, nous avons procédé à des expérimentations intensives sur des bases de test provenant de l'UCI Irvine (Blake 1998) ainsi que sur des jeux d'essai synthétiques. Nous avons pris en compte plusieurs critères d'évaluation, principalement la performance prédictive, la robustesse (dégradation de la performance entre apprentissage et test), la taille des discrétisations, le comportement vis à vis des attributs bruités et le temps de calcul des groupages. Une analyse multicritères des résultats permet de comparer en détail les différentes méthodes. Cette évaluation approfondie confirme les résultats théoriques et montre que la méthode MODL est la plus performante sur l'ensemble des critères considérés.

Le reste du document est organisé de la façon suivante. La partie 2 introduit la nouvelle méthode de groupage MODL en détaillant son critère d'évaluation, en étudiant différents algorithmes d'optimisation, enfin en introduisant la gestion du groupe poubelle. La partie 3 procède à des expérimentations comparatives permettant une évaluation multicritères des méthodes de groupage.

2 Groupage MODL

Dans cette partie, on présente dans un premier temps une formalisation du problème du groupage, permettant d'aboutir à un critère d'évaluation caractérisant les groupages optimaux au sens de Bayes. Dans un second temps, on propose des algorithmes d'optimisation permettant de rechercher les groupages optimaux, de façon exacte ou approchée. Enfin, on présente la gestion du groupe poubelle et son intégration dans le critère d'évaluation et dans les algorithmes d'optimisation.

2.1 Critère d'évaluation

On présente ici une formulation du problème du groupage puis une liste de résultats théoriques, dont les démonstrations sont données en annexe. On introduit notamment le critère d'évaluation des groupages permettant de caractériser la solution optimale au sens de Bayes.

2.1.1 Présentation

Tout d'abord, afin d'illustrer la problématique du groupage, nous présentons dans la table 1 un exemple extrait de la base Mushroom de l'UCI (Blake and Merz 1998). Il s'agit de prédire le caractère comestible (EDIBLE) ou non (POISONOUS) des champignons en fonction de leur caractéristique physique comme par exemple la couleur du chapeau (CapColor). La table de contingence du tableau 1 résume le fichier de données en mémorisant pour chaque couleur de chapeau la proportion des champignons comestibles ou non.

Tableau 1 : Table de contingence pour l'attribut CapColor du jeu de donnée Mushroom de l'UCI

	EDIBLE	POISONOUS	Frequency
BROWN	55.2%	44.8%	1610
GRAY	61.2%	38.8%	1458
RED	40.2%	59.8%	1066
YELLOW	38.4%	61.6%	743
WHITE	69.9%	30.1%	711
BUFF	30.3%	69.7%	122
PINK	39.6%	60.4%	101
CINNAMON	71.0%	29.0%	31
GREEN	100.0%	0.0%	13
PURPLE	100.0%	0.0%	10

Le groupage de l'attribut consiste à partitionner les valeurs de façon à réduire leur nombre tout en conservant au maximum l'information sur la classe à prédire. Il s'agit donc d'un compromis entre valeur statistique (effectifs par groupe suffisants pour fiabiliser la prédiction) et valeur informationnelle (proportions différentes par groupe pour améliorer la finesse de la prédiction). La figure 1 fournit un exemple de groupage de l'attribut CapColor de Mushroom (réalisé avec la méthode MODL). On voit dans cet exemple que BROWN et GRAY constituent chacun un groupe indépendant en raison de leur effectif très important, bien que la différence de proportion des champignons comestibles soit faible entre les deux groupes. En revanche, BUFF (avec 30.3% de champignons comestibles) a été intégré dans le groupe {RED, YELLOW, BUFF, PINK} (de proportion avoisinant 40%), son effectif relativement faible ne justifiant pas la création d'un groupe à part. A l'autre extrême on trouve les modalités GREEN et PURPLE qui ont donné lieu à un seul groupe en dépit de leur très faible effectif. Cet exemple didactique illustre la complexité du compromis à trouver entre valeur statistique et informationnelle. Un autre aspect du problème est l'aspect combinatoire: le nombre de groupages potentiels croît de façon exponentielle avec le nombre de modalités.

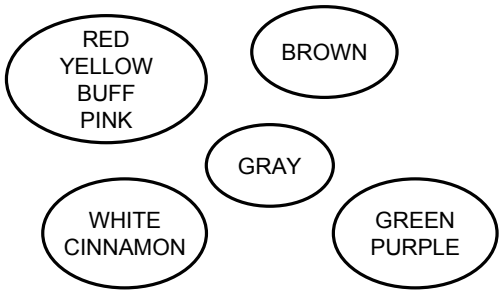

Figure 1 : Un exemple de groupage des valeurs de l'attribut CapColor de la base Mushroom

2.1.2 Formulation du groupage

Chaque instance de donnée est décrite par une valeur descriptive et une classe (valeur cible à prédire). Les valeurs descriptives sont symboliques, ce qui signifie que l'on peut les distinguer deux à deux, mais qu'on ne peut pas les ordonner de façon "naturelle".

Les données en entrée constituent une chaîne d'instances S (valeur descriptive, classe). On suppose que le nombre d'instances, le nombre de classes et les valeurs descriptives des instances sont connues à l'avance. On connaît donc le nombre de valeurs descriptives différentes, ainsi que leur fréquence dans la chaîne d'instances.

Définition: Un modèle de groupage est dit *standard* s'il respecte les conditions suivantes:
- il permet de définir une partition des valeurs descriptives en groupes,
- la distribution des classes sur chaque groupe est définie uniquement par l'effectif des classes dans ce groupe.

On dira qu'un tel modèle de groupage est de type SGM (Standard Grouping Model).

Notations:
n: nombre d'instance de la chaîne à grouper,
J: nombre de classes,
I: nombre de valeurs descriptives,
n_i: nombre d'instances pour la valeur descriptive i,

n_{ij}: nombre d'instances de la classe *j* pour la valeur descriptive *i*,
K: nombre de groupes,
k(i): index du groupe auquel est rattaché la valeur descriptive *i*,
n_k: nombre d'instances pour le groupe *k*,
n_{kj}: nombre d'instances de classe *j* pour le groupe *k*.

Les données connues à l'avance sont *n*, *J*, *I* et n_i.
Un groupage de type SGM est entièrement caractérisé par le choix de *K*, des *k(i)* et des n_{kj}.

Définition: Un modèle SGM est *compatible* avec une chaîne d'instances si les sous-ensembles d'instances correspondant aux groupes définis par le modèle ont une distribution de classe identique à celle définie par le modèle.

Théorème: Un modèle SGM d'une chaîne d'instances ne peut-être optimal au sens de Bayes que s'il est compatible avec cette chaîne.

En effet, la probabilité qu'une chaîne *S* non compatible avec un modèle SGM soit conforme à ce modèle est par définition nulle. L'intérêt de ce résultat est que tout algorithme d'optimisation d'un groupage SGM d'une chaîne *S* peut se contenter de parcourir les modèles compatibles avec *S*, c'est à dire que l'on peut se limiter au choix d'une partition des valeurs descriptives en groupes, le choix des distributions par groupe étant déduit des caractéristiques de la chaîne *S*.

Définition: On appelle *a priori d'un modèle de groupage* toute distribution de probabilité portant sur les réalisations possibles du modèle.

2.1.3 Groupage MODL standard

Définition: On appelle *a priori à trois étages* l'a priori de modèle SGM basé sur les hypothèses suivantes:
- le nombre *K* de groupes est compris entre 1 et *I*, de façon équiprobable,
- pour un nombre de groupes donné, toutes les partitions des *I* valeurs initiales en *K* groupes sont équiprobables,
- pour un groupe donné, toutes les distributions de classes sont équiprobables,
- les distributions des classes sur chaque groupe sont indépendantes les unes des autres.

Cet a priori est essentiellement un a priori uniforme à chaque niveau de la hiérarchie des paramètres des modèles SGM. L'hypothèse supplémentaire d'indépendance des groupes est généralement admise. La plupart des méthodes de groupage existantes cherchent à s'en approcher, au moins implicitement. Par exemple, la méthode CHAID basée sur le critère statistique du Khi2 recherche des groupes qui soient le plus indépendant possible deux à deux.

Théorème: Un modèle de groupage standard *M* suivant un a priori à trois étages est un modèle de prédiction optimal au sens de Bayes si son évaluation par la formule suivante est minimale sur l'ensemble de tous les modèles:

$$Value(M) = \log(I) + \log(B(I,K)) + \sum_{k=1}^{K} \log\left(C_{n_k+J-1}^{J-1}\right) + \sum_{k=1}^{K} \log\left(n_k! / n_{k,1}! n_{k,2}! .. n_{k,J}!\right).$$

$B(I,K)$ est le nombre de partitions de *I* valeurs en *K* groupes (éventuellement vides). Les formules de calcul de ce nombre, baptisé nombre de Bell "généralisé", sont détaillées en annexe.
Le premier terme de la formule correspond au choix du nombre de groupes et le second terme au choix de la partition des valeurs descriptives en groupes. Le troisième terme représente le choix des distributions de classes cibles dans chaque groupe et le dernier terme le codage des données connaissant le modèle de groupage.

Théorème: Dans un modèle SGM optimal suivant un a priori à trois étages, deux valeurs descriptives mono-classe et de même classe sont nécessairement dans le même groupe.

Cette propriété correspond au caractère "well-behaved" d'une méthode de groupage tel qu'il est défini dans (Elomaa & Rousu 1997). L'intérêt de ce théorème est essentiellement de valider l'a priori à trois étages, en vérifiant qu'une propriété "intuitive" des groupages optimaux découle de l'a priori. Un autre intérêt de ce théorème est de permettre d'améliorer le temps de calcul des algorithmes d'optimisation des groupages, en fusionnant dans un prétraitement toutes les valeurs descriptives pures selon leur classe cible, ce qui permet de diminuer ainsi le nombre de valeurs à traiter.

Théorème: Dans un modèle SGM suivant un a priori à trois étages dans le cas de deux classes cibles, s'il y a autant de valeurs descriptives que d'instances, alors le groupage optimal est réduit à un seul groupe contenant toutes les valeurs

descriptives.

Il s'agit cette fois encore de valider l'a priori a trois étages, en vérifiant rigoureusement une autre propriété "intuitive", à savoir qu'il n'est pas possible de généraliser à partir d'expériences en un seul exemplaire (une seule instance par valeur descriptive).

Conjecture: Dans un modèle SGM suivant un a priori à trois étages et dans le cas de deux classes cibles, on peut ordonner les valeurs descriptives par proportion croissante de la première classe cible. Alors, toute valeur descriptive de proportion comprise entre deux valeurs descriptives incluse dans un même groupe doit nécessairement appartenir à ce groupe dans le groupage optimal.

Cette conjecture a été démontrée pour d'autres critères de partitionnement (critère de Gini (Breiman, 1984) et de Kolmogoriv-Smirnov (Asseraf, 1996)) et vérifiée expérimentalement dans un large domaine de valeurs (voir annexe). Cette propriété est très intéressante, car elle permet de diminuer la complexité des algorithmes de recherche des partitions. Cette conjecture n'a pas été démontrée analytiquement. Elle sera néanmoins considérée comme vraie par la suite en raison de son intérêt pour l'optimisation des algorithmes.

2.1.4 Commentaires

Dans l'a priori à trois étages, on suppose que les partitions de I valeurs en K groupes peuvent éventuellement donner lieu à des groupes vides. Cela ne paraît pas absurde, dans la mesure où le nombre de groupe étant fixé, le choix d'un groupe pour chaque valeur peut naturellement conduire à laisser des groupes vides, notamment quand le nombre de groupes est proche du nombre de valeurs à grouper. Une variante consiste à imposer des partitions en K groupes non vides. Le nombre de partitions des I valeurs en K groupes est alors égal au nombre de Stirling de seconde espèce $S(I,K)$, au lieu du "nombre de Bell généralisé" $B(I,K)$. On a la relation suivante:

$$B(n,k)=\sum_{i=1}^{k} S(n,i)$$

On peut vérifier que pour les faibles nombres de groupes ($K < I/\ln(I)$), les nombres $S(I,K)$ et $B(I,K)$ sont très proches, mais qu'au delà de ce seuil, les $B(I,K)$ continuent de croître très légèrement, alors que les $S(I,K)$ se mettent à décroître jusqu'à atteindre la valeur 1 pour $K=I$. Le choix de la variante d'a priori avec les groupes non vides peut se justifier en remarquant que la longueur de codage correspondante est systématiquement plus faible qu'avec le critère initial. Si l'on se réfère à la théorie de la complexité algorithmique de Kolmogorov, un critère donnant lieu à un codage plus compact paraît plus proche de l'optimum théorique de l'algorithme de longueur minimale et semble donc préférable. Néanmoins, on montre dans ce cas (voir annexe) que les théorèmes sur les propriétés "intuitives" des groupages ne sont plus valides si l'on évalue le nombre de partitions par les nombres de Stirling de seconde espèce. Principalement, dans le cas où il n'y a qu'une instance par valeur descriptive, le groupage optimal peut dans certains cas impliquer autant de groupes que d'instances (bien que cette solution soit de coût très proche de celui du groupage en un seul groupe). Ces résultats rendent la variante moins intéressante que la version originale de l'a priori.

En fait, la théorie de la complexité de Kolmogorov et l'approche MDL qui s'en inspire ont une validité asymptotique, alors que l'approche bayésienne de MODL est optimale y compris pour des jeux de données de très petite taille. De plus, l'approche MDL est une heuristique guidée par l'approche optimale de Kolmogorov, qui elle n'est pas calculable. Ceci explique que pour l'approche MODL (optimale pour un a priori donné), il n'est pas possible de guider le choix de l'a priori uniquement par la minimisation de la longueur de codage résultant. La validation du choix de l'a priori par la vérification de certaines propriétés "intuitives" est une méthodologie délicate, mais qui semble plus pertinente.

2.2 Algorithmes d'optimisation

On présente dans cette partie différents algorithmes d'optimisation pour la recherche d'un groupage de coût minimal. On distingue le cas à deux classes cibles permettant de réutiliser les algorithmes de discrétisation MODL et le cas général, où l'on propose un algorithme glouton ascendant d'agrégation itérative des valeurs descriptives, puis des améliorations sous forme de pré-optimisation et post-optimisation.

2.2.1 Cas à deux classes cibles

Si l'on admet la conjecture sur la compatibilité du groupage optimal avec l'ordre des valeurs descriptives triée par fréquence cible croissante, on peut alors utiliser les mêmes algorithmes que pour la méthode de discrétisation MODL (Boullé 2004). On obtient un algorithme optimal en $O(n^3)$ (où plus précisément en $O(n.\log(n)+I^3)$ en se basant sur un algorithme de programmation dynamique. On peut également utiliser l'heuristique de discrétisation avec post-optimisation poussée, dont la complexité algorithmique est $O(n.\log(n))$.

2.2.2 Cas général

Le principe de l'algorithme glouton consiste à initialiser la solution avec autant de groupes que de valeurs descriptives, à rechercher la meilleure fusion de groupes, et à réitérer ce processus tant qu'il y a amélioration du critère de groupage. Une implémentation naïve de cet algorithme conduit à une complexité algorithmique en $O(n^3)$. En utilisant l'additivité du critère MODL par rapport aux groupes, en bufferisant les résultats d'évaluation des fusions de groupes, et en mémorisant l'ordre de préférence des fusions dans une liste triée maintenable (de type AVL Binary Search Tree par exemple), on peut ramener cette complexité à $O(n^2.\log(n))$.

Algorithme

- Initialisation
  - Tri des valeurs de l'attribut descriptif : en $O(n.\log(n))$
  - Création d'un groupe élémentaire par valeur de la loi source : en $O(n)$
  - Calcul des $GroupCost$ initiaux : en $O(n)$
  - Calcul des $\Delta GroupCost$ suite à chaque fusion de groupes: en $O(n^2)$
  - Tri des fusions de groupes par valeur de $\Delta GroupCost$: en $O(n^2.\log(n))$
- Optimisation du groupage
  - Répéter: n étapes
    - Chercher la meilleure fusion : en O(1) en prenant le premier élément de la liste triée
    - Evaluer la condition d'arrêt
      - ✓ Arrêter si $\Delta GroupingCost = \Delta PartitionCost + \Delta GroupCost \geq 0$
      - ✓ Continuer sinon
    - Si continuer : effectuer la fusion de groupe
      - Calcul du $GroupCost$ pour le nouveau groupe : en O(1)
      - Calcul des $\Delta GroupCost$ pour les groupes fusionnables avec le nouveau groupe : en $O(n)$
      - Mise à jour de la liste triée des $\Delta GroupCost$: en $O(n.\log(n))$
        - ✓ Suppression du $\Delta GroupCost$ ayant conduit au nouveau groupe
        - ✓ Suppression des anciens $\Delta GroupCost$ des groupes fusionnables avec les deux sous-groupes du nouveau groupe
        - ✓ Ajout des nouveaux $\Delta GroupCost$ des groupes fusionnables avec le nouveau groupe

On peut noter que l'occupation mémoire nécessaire pour l'algorithme est également en $O(n^2.\log(n))$. On doit en effet mémoriser n $GroupCost$, n^2 $\Delta GroupCost$, et une structure de liste triée de type arbre binaire de recherche équilibré qui a une occupation mémoire de $O(n^2.\log(n))$.

Notons que cette complexité algorithmique correspond au pire des cas où le nombre de valeurs descriptives I est égal au nombre d'instances n. De façon plus précise, la complexité algorithmique est $O(n.\log(n)+I^2.\log(I))$.

2.2.3 Pré-optimisation

Dans le cas général, la complexité algorithmique $O(n.\log(n)+I^2.\log(I))$ peut devenir trop grande. Afin de contrôler cette complexité et de la ramener systématiquement à $O(n.\log(n))$, il suffit dans un prétraitement de ramener le nombre de modalités initiales I à un nombre $I' \leq \sqrt{n}$.

Regroupement des valeurs pures

L'objectif est ici de regrouper par classe cible les valeurs descriptives initiales "pures" (associées à une unique classe cible). Ce prétraitement est compatible avec la solution optimale.

Identification d'un nombre restreint de groupes initiaux à partir de groupages partiels par classe cible

Dans une première étape, on effectue un groupage optimisé pour chaque classe cible, en mode une classe contre toutes les autres (utilisation des algorithmes spécifiques à deux classes cibles). La complexité de cette étape est $O(J.n.\log(n))$. Dans une seconde étape, on identifie les sous-groupes stables pour l'ensemble des J groupages ainsi constitués. Pour cela, il suffit d'associer à chaque modalité une clé de groupage résultant de la concaténation de ses index de groupes sur chacun des J axes de groupages. Cette clé sert de clé de hachage et permet d'identifier les sous-groupes stables sur tous les axes. Cette seconde étape a ainsi une complexité algorithmique en $O(I.J)$.

Groupage majoritaire des valeurs descriptives peu fréquentes

On crée au plus un groupe par classe cible pour agglomérer les valeurs descriptives surnuméraires ayant même classe cible majoritaire. Les valeurs descriptives les moins fréquentes sont ainsi groupées jusqu'à obtenir le nombre de valeurs désirées. Cela ne nécessite qu'un tri des valeurs par effectif croissant, en $O(I.\log(I))$. Cette étape, indispensable pour contrôler la complexité algorithmique de l'heuristique, peut entraver la qualité de la solution. La post-optimisation décrite ci-dessous est

alors nécessaire pour raffiner la solution trouvée.

Il est à noter que cette étape, utile pour borner théoriquement la complexité algorithmique du groupage, est très rarement activée en pratique. En effet, dans les cas des très grand nombres de modalités, les modalités de très faible effectif sont fréquemment pures et ainsi regroupées inconditionnellement par la première étape de prétraitement. Pour la deuxième étape de prétraitement, le coût de la partie "partition" du groupage est prohibitif dans le critère d'évaluation des groupages, ce qui favorise les groupages partiels ayant très peu de groupes, induisant un nombre limité de groupes initiaux.

2.2.4 Post-optimisation

L'algorithme glouton du cas général court le risque de tomber dans un optimum local, ce qui justifie une post-optimisation.

Exhaustive Merge

Cette heuristique consiste à continuer à forcer les merges jusqu'à obtenir un seul groupe final, puis à retenir le meilleur groupage rencontré. Avec les garanties du prétraitement, on reste en O(n.log(n)).

Optimisation locale

A nombre de groupes constant, il s'agit de déplacer les valeurs d'un groupe à un autre. Pour chaque valeur, on évalue la variation de coût entraînée par son transfert vers un autre groupe. On effectue ces transferts tant qu'il y a amélioration du critère d'évaluation MODL du groupage. En fait, chaque valeur descriptive est ainsi attirée vers son groupe le plus proche. A la manière des K-moyennes, cet algorithme converge très rapidement, sans que l'on puisse le démontrer. Ce nombre d'étape est d'autant plus faible que l'on part d'une solution déjà optimisée (sauf si l'on a prétraité les valeurs rares dans la pré-optimisation). Chaque étape d'amélioration est en O($I.K$). On n'utilise cette heuristique que si $I.K<=n.log(n)$.

Merges optimisés

Une seconde heuristique consiste à rechercher un nouveau groupage en supprimant un groupe. L'heuristique consiste dans un premier temps à rechercher la meilleure fusion de groupes, à forcer cette fusion inconditionnellement, puis à post-optimiser le groupage au moyen de l'optimisation locale, par échange de valeurs entre les groupes. Le nouveau groupage est accepté s'il y a amélioration du critère et dans ce cas, l'heuristique est réitérée. Chaque étape d'amélioration est en O($K^2 + I.K$), avec les mêmes conditions d'application que pour l'heuristique précédente. Une variante consiste à réitérer l'heuristique en bornant par une constante (typiquement 2 ou 3) le nombre d'étapes autorisées sans amélioration.

2.3 Prise en compte du groupe poubelle

Dans le cas d'un très grand nombre de valeurs descriptives, le nombre de possibilités de partitions des valeurs en groupes est tel qu'il implique un risque de sur-apprentissage important. Dans ce cas, une méthode statistique fiable se prononcera fréquemment en faveur d'un seul groupe, alors qu'il existe potentiellement quelques valeurs significativement représentées. Afin de repousser les limites de la recherche d'information dans un attribut symbolique, on introduit un groupe poubelle regroupant les valeurs descriptives de faible effectif. On permet ainsi au groupage de travailler sur un nombre "raisonnable" de valeurs restant à grouper. Le problème est d'ajuster l'effectif minimum des valeurs échappant au groupe poubelle. Il s'agit de trouver un compromis: un effectif important permettra de proposer un groupage utile et fiable, au détriment d'une perte informationnelle due au mélange des valeurs dans le groupe poubelle.

On propose une extension du groupage MODL standard présenté précédemment, permettant d'incorporer le groupe poubelle à la fois dans le critère d'évaluation des groupages et dans les algorithmes d'optimisation.

2.3.1 Groupage MODL basique

On cherche ici un modèle élémentaire, où les valeurs descriptives de faible effectif sont groupées dans le groupe poubelle et les autres valeurs sont laissées telles quelles (sans groupage). Il s'agit ici de déterminer l'effectif minimum des valeurs échappant au groupe poubelle, en définissant soit un a priori sur la distribution de ces effectifs minimum, soit un a priori sur la distribution des nombres de valeurs restantes après le prétraitement de poubellisation. On va privilégier les faibles nombres de groupe et utiliser à cet effet l'a priori universel sur les entiers (Rissanen 1983) détaillé en annexe.

Soit F l'effectif minimum des valeurs échappant à la poubelle et $I(F)$ le nombre de valeurs restantes (groupe poubelle compris) après fusion des valeurs de faible effectif dans la poubelle.

Définition: On appelle *a priori basique avec poubelle* l'a priori de modèle SGM basé sur les hypothèses suivantes:

- le nombre de valeurs restantes après prétraitement de poubellisation suit l'a priori universel sur les entiers,
- chaque valeur restante après la poubellisation constitue un groupe,
- pour un groupe donné, toutes les distributions de classes sont équiprobables,
- les distributions des classes sur chaque groupe sont indépendantes les unes des autres.

On définit alors la variante du critère MODL prenant en compte le groupe poubelle dans le cas basique sans groupage des valeurs restantes.

Théorème: Un modèle de groupage standard *M* suivant un a priori basique avec poubelle est un modèle de prédiction optimal au sens de Bayes si son évaluation par la formule suivante est minimale sur l'ensemble de tous les modèles:

$$Value(M) = L(K)\log(2) + \sum_{k=1}^{K} \log\left(C_{n_k+J-1}^{J-1}\right) + \sum_{k=1}^{K} \log\left(n_k!/n_{k,1}!n_{k,2}!..n_{k,J}!\right)$$

Preuve:

Le terme $L(K)$ correspond à l'a priori universel sur les entiers, normalisé par un facteur log(2) pour utiliser la même base de log que dans le reste du critère. Le nombre de groupe K étant fixé égal à $I(F)$ une fois F fixé, il n'y pas ici de coût de partitionnement des valeurs en groupes. Le reste du critère est identique au cas du groupage MODL standard, dont la preuve est fournie en annexe.

Algorithme:

L'algorithme se limite ici au tri des valeurs descriptives par fréquence croissante (en $O(I.\log(I))$), puis au test de toutes les regroupements possibles sur cette liste en partant des modalités les plus rares (en $O(I)$) pour choisir le seuil d'effectif minimum minimisant le critère.

2.3.2 Groupage MODL

La version générale du groupage MODL est une extension du groupage MODL standard avec prise en compte du groupe poubelle. Dans un premier temps, il faut choisir si l'on va utiliser ou non une poubelle, puis si nécessaire définir l'effectif minimum des valeurs descriptives échappant à la poubelle. A l'issue de ce prétraitement, on procède au groupage standard des valeurs restantes en les partitionnant en groupes. On va ici considérer le groupage MODL général comme une variante du groupage MODL standard dont on ne veut s'éloigner que si nécessaire. Dans le cas avec poubelle, on fait alors porter l'a priori sur l'effectif minimum des valeurs échappant au groupe poubelle en privilégiant les petits effectifs (proches du cas sans poubelle).

Définition: On appelle *a priori à trois étages avec poubelle* l'a priori de modèle SGM basé sur les hypothèses suivantes:

- utiliser ou non une poubelle sont deux choix équiprobables,
- l'effectif minimum F des valeurs initiales échappant à la poubelle suit l'a priori universel sur les entiers,
- le nombre K de groupes est compris entre 1 et $I(F)$, de façon équiprobable,
- pour un nombre de groupes donné, toutes les partitions des $I(F)$ valeurs restantes en K groupes sont équiprobables,
- pour un groupe donné, toutes les distributions de classes sont équiprobables,
- les distributions des classes sur chaque groupe sont indépendantes les unes des autres.

La variante du critère MODL prenant en compte le groupe poubelle dans le cas à trois étages est définie par le théorème suivant, en notant $1_{[2,+\infty[}(x)$ la fonction indicatrice de l'intervalle $[2,+\infty[$ valant 0 avant la borne de l'intervalle et 1 après.

Théorème: Un modèle de groupage standard *M* suivant un a priori à trois étages avec poubelle est un modèle de prédiction optimal au sens de Bayes si son évaluation par la formule suivante est minimale sur l'ensemble de tous les modèles:

$$Value(M) = \log(2) + 1_{[2,+\infty[}(F)L(F)\log(2) + \log(I(F)) + \log(B(I(F),K)) + \sum_{k=1}^{K} \log\left(C_{n_k+J-1}^{J-1}\right) + \sum_{k=1}^{K} \log\left(n_k!/n_{k,1}!n_{k,2}!..n_{k,J}!\right)$$

Preuve:

Le terme log(2) encode le choix d'utiliser ou non une poubelle. Dans ce dernier cas uniquement (si F>1), il faut encoder l'effectif minimum des valeurs hors poubelle, en utilisant l'a priori universel sur les entiers. On obtient un nombre de modalités restantes $I(F)$, permettant de coder la suite du critère de manière identique au cas du groupage MODL standard, dont la preuve est fournie en annexe.

En pratique, l'amélioration des groupages est assez faible pour un surcoût algorithmique potentiellement important. Néanmoins, dans les cas de nombreuses valeurs descriptives, l'amélioration de la prédiction en apprentissage n'est plus négligeable et la robustesse (mesurée par l'écart entre les performances en test et apprentissage) est significativement améliorée. Il est alors intéressant d'envisager de prendre en compte ce critère si l'on peut trouver une heuristique de coût limité à $O(n.\log(n))$.

Algorithme:

On propose un algorithme glouton en partant d'une solution initiale de groupage optimisé sans contrainte d'effectif minimum. On se ramène ainsi au groupage standard dans cette première étape. On évalue ensuite les solutions avec poubelle

par taille croissante de l'effectif minimum des valeurs hors poubelle. Pour chaque taille, on effectue d'abord le prétraitement de poubellisation (O(n)), puis on réutilise l'algorithme de groupage MODL standard. On continue tant qu'il y a amélioration du critère. Dans une variante, on continue tant qu'il y a eu au moins une amélioration dans les T dernières (par exemple T=3) évaluations.

Dans le pire des cas, il semble que n évaluations de tailles de poubelle soient potentiellement nécessaires. En fait, il ne peut y avoir au plus que O($\sqrt{n}$) effectifs différents dans un ensemble de I valeurs d'effectif total égal à n. En effet, si on note n_i l'effectif de la valeur descriptive i et si on suppose que tous les effectifs sont distincts, le nombre total de valeurs d'effectif est borné par la contrainte suivante:

$$\sum_{i=1}^{I} n_i = n .$$

Dans ce cas, I est maximum quand les n_i sont minimaux et suivent une progression arithmétique:

$$I(I+1)/2 = n .$$

Il y a bien au pire O($\sqrt{n}$) effectifs minimum distincts à évaluer, CQFD.

La complexité algorithmique du groupage MODL général est alors O($n\sqrt{n}$ log(n)) dans le pire des cas. En pratique, le coût d'encodage des effectifs minimums étant négligeable devant les coûts d'encodage de la partition et de la distribution des instances, le gain du groupe poubelle ne se réalise que dans les cas où une faible variation de l'effectif minimum entraîne une grande diminution du nombre de valeurs descriptives restant à grouper. Cette caractéristique entraîne un arrêt très rapide en moyenne de l'algorithme glouton, dont la complexité empirique est alors O($T.n$.log(n)). Des expérimentations intensives ont montré que la qualité des groupages produits par l'heuristique gloutonne était quasiment identique à celle d'un algorithme évaluant tous les effectifs minimaux possibles.

3 Evaluation

3.1 Critères d'évaluation du groupage

L'objectif du groupage est de trouver un compromis entre valeur informationnelle et valeur statistique, de diminuer le nombre de groupes tout en conservant au maximum l'information contenue dans les modalités initiales. Les méthodes de prétraitement des attributs, discrétisation et groupage, sont classiquement évaluées au moyen d'un algorithme inductif, en général un arbre de décision ou un prédicteur Bayesien Naïf. Le taux d'erreur en prédiction est comparé avec ou sans prétraitement des attributs sur une dizaine de bases de l'UCI (Blake and Merz 1998). Ce type d'évaluation est insuffisant et ne permet pas de dégager l'apport intrinsèque de chaque méthode, celui-ci étant perturbé par l'algorithme inductif utilisé en aval. Par ailleurs, la mesure du seul taux d'erreur est insuffisante. Les applications basées sur le scoring des instances par exemple nécessitent d'évaluer la probabilité prédite pour chaque classe cible afin d'ordonner les instances par score décroissant. Enfin, dans le cas spécifique du groupage, l'étude des groupes constitués est intrinsèquement intéressante dans une phase exploratoire du data mining et peut fournir des informations pertinentes pour l'explication de l'apport des attributs.

Dans le cas du groupage, l'enjeu est de réduire au maximum le nombre de valeurs descriptives initiales tout en assurant une perte d'information minimale. Le prédicteur optimal est le prédicteur de Bayes, ce qui dans le cas d'un prédicteur univarié basé sur un attribut symbolique signifie que la prédiction optimale est atteinte en prédisant pour chaque valeur descriptive la classe majoritaire observée en apprentissage. Le groupage optimum vis à vis de la mesure du taux d'erreur consiste donc à ne rien faire. La recherche du compromis de groupage peut donc se reformuler sous la forme d'un problème bi-critères, pour lequel il est facile de situer les cas extrêmes, à savoir d'une part le prédicteur bayésien qui minimise la perte d'information mais ne procède à aucun groupage, d'autre part le prédicteur majoritaire, qui prédit systématiquement la classe majoritaire de l'échantillon en groupant toutes les valeurs descriptives dans un seul groupe.

Le nombre de groupes est une bonne mesure de la taille du groupage. Le taux d'erreur de prédiction est par contre peu pertinent pour mesurer la perte d'information due au groupage. Ce taux d'erreur est un résumé très synthétique de la performance d'une méthode inductive. Il ne prend en compte que la classe majoritaire et ne permet pas de différencier finement les méthodes entre elles. Ainsi, pour un problème où la classe majoritaire est majoritaire dans toutes les valeurs descriptives, le prédicteur majoritaire obtient la même performance que le prédicteur optimal de bayes, alors qu'il est incapable de distinguer les individus prédits correctement avec une probabilité proche de 100% (modalités pures) des individus prédits presque au hasard avec une probabilité légèrement supérieure à 50%.

La divergence de Kullback-Leibler est une mesure de différence entre deux distributions de probabilités, qui tend vers 0 quand les deux distributions convergent. Cette mesure parait adaptée pour évaluer la perte d'information entre la distribution de probabilités des classes cibles estimée à partir des groupes constitués en apprentissage et la distribution des probabilités des classes cibles observée à partir des valeurs descriptives en test. Pour une valeur descriptive donnée, soit p_j la probabilité de la classe cible j, observée sur l'ensemble de test en se basant sur toutes les instances associées à la valeur descriptive, et soit q_j la probabilité de la classe cible j, estimée sur l'ensemble d'apprentissage en se basant sur toutes les instances du groupe contenant la valeur descriptive. La divergence de Kullback-Leibler entre la distribution observée en test et la distribution estimée en apprentissage est définie par:

$$D(p \,||\, q) = \sum_{j=1}^{J} p_j \log \frac{p_j}{q_j}$$

Les probabilités q_j sont estimées par l'estimateur de Laplace afin de gérer les probabilités nulles, ce qui n'est pas nécessaire pour les probabilités p_j. La divergence globale est calculée en prenant la moyenne des divergences sur l'ensemble des instances de l'échantillon de test. Afin de permettre les comparaisons entre méthodes sur des jeux d'essai de nature et de taille différentes, on normalise les mesures de Kullback-Leibler précédemment définies par le résultat obtenu avec la méthode consistant à ne rien grouper. Cette dernière méthode obtient ainsi une valeur plancher de 1, dont les autres méthodes doivent s'approcher le plus possible. Les moyennes synthétiques résumant les résultats sur un grand nombre d'expérimentations sont des moyennes géométriques dans le cas de la divergence de Kullback-Leibler.

En définitive, les critères utilisés dans l'expérimentation sont le taux de bonne prédiction (en test, mais également en apprentissage afin d'évaluer la robustesse des méthodes), le nombre de groupes produits et la divergence de Kullback-Leibler.

3.2 Méthodes évaluées

Nous allons évaluer la méthode MODL en la comparant avec d'autres méthodes basées sur le critère du Khi2 ainsi qu'avec la méthode basée sur le Gain Ratio popularisée par son utilisation dans l'algorithme C4.5. Toutes ces méthodes utilisent un algorithme glouton de fusions itérative des modalités initiales.

3.2.1 Méthode MODL

Il s'agit de la méthode décrite dans ce papier. La méthode MODL utilisé est la méthode générale avec prise en compte d'un groupe poubelle.

Des variantes de la méthode sont également évaluées pour des raisons comparatives. Il s'agit de la méthode MODLStandard (sans prise en compte du groupe poubelle), de la méthode BasicMODL (prise en compte uniquement du groupe poubelle, sans groupage des modalités restantes) et de la méthode ExhaustiveMODL (MODL en forçant les groupages à contenir au plus deux groupes).

3.2.2 Méthode Khiops

La méthode Khiops (Boullé 2003) minimise la probabilité d'indépendance entre l'attribut groupé et l'attribut à prédire, en s'assurant en plus que les variations du Khi2 observées lors des regroupements sont significativement différentes de celles provenant du groupage d'un attribut descriptif indépendant de l'attribut cible. On utilise le seuil de 95% pour ce test de significativité.

En prétraitement, toutes les valeurs descriptives n'atteignant pas un effectif minimum sont regroupées dans une modalité spéciale. Cet effectif minimum est calculé au plus juste pour permettre un compromis entre fiabilité du test du Khi2 et finesse de la partition produite.

3.2.3 Méthode Tschuprow

Les coefficients de contingence de Pearson, de Cramer et de Tschuprow sont basés sur une normalisation à 1 de la valeur du Khi2, ce qui les rend moins dépendants de la taille du tableau de contingence que la valeur du Khi2 initiale. Ces coefficients sont définis de la façon suivante.

Coefficient de contingence de Pearson: $C = \sqrt{\frac{Khi2}{n + Khi2}}$

Coefficient de Cramer: $v = \sqrt{\frac{Khi2}{n \min(I-1, J-1)}}$

Coefficient de Tschuprow: $t = \sqrt{\frac{Khi2}{n\sqrt{(I-1)(J-1)}}}$

On peut montrer qu'en dépit de leur normalisation à 1, ces coefficients ne constituent pas une évaluation des tableaux de contingence équitable vis à vis du nombre de lignes, spécialement quand ils sont utilisés dans le cadre d'un algorithme de regroupement des lignes d'un tableau de contingence. Ainsi, les numérateurs des coefficients de contingence de Cramer et de Tschuprow ne peuvent que décroître suite à une fusion de deux lignes (propriétés du DeltaKhi2). Le dénominateur du coefficient de contingence de Pearson décroît proportionnellement moins vite que son numérateur, ce qui fait que ce coefficient ne peut que décroître. Le coefficient de contingence de Pearson est donc inutilisable pour le problème du groupage, car il est maximal quand aucun groupage n'est effectué. En ce qui concerne le coefficient de Cramer, dans le cas standard où le nombre de lignes est supérieur au nombre de colonnes, le coefficient ne peut que décroître (son dénominateur ne variant pas). Le coefficient de Cramer est donc également inutilisable pour le problème du groupage. Le cas du Tschuprow est plus subtil. Contrairement au Cramer, le Tschuprow ne peut atteindre sa borne de 1 que dans le cas où le nombre de lignes est égal au nombre de colonnes et sa borne théorique est d'autant plus proche de 1 que le tableau de contingence est proche d'un tableau

carré. En conséquence, cet effet à tendance à favoriser les partitions ayant même nombre de groupes que de classes cibles. Cela reste à vérifier dans les expérimentations. Par ailleurs, la méthode basée sur le Tschuprow ne peut produire que des partitions ayant au moins deux groupes.

3.2.4 Méthode CHAID

La méthode CHAID (Kass 1980) applique le critère du Khi2 non pas globalement à la partition comme dans les méthodes précédentes, mais localement à deux groupes dont la fusion est évaluée. Les groupes sont fusionnés s'ils sont statistiquement similaires. On utilisera le seuil de 95% pour le test d'arrêt. CHAID envisage également de remettre en question des fusions de modalités en éclatant les groupes constitués. Selon l'auteur lui-même, cette particularité de l'algorithme est en pratique rarement utile. On n'utilisera pas cette extension pour les expérimentations. Il est à noter que dans le cas où il n'y a que deux valeurs descriptives initiales, la méthode CHAID est identique à la méthode Khiops (exceptée la gestion de l'effectif minimum par groupe dans une phase de prétraitement pour Khiops).

La variante ExhaustiveCHAID (popularisée par son implémentation dans le produit AnswerTree de SPSS) est également évaluée. Cela permet une comparaison complémentaire entre MODL et CHAID, le nombre de groupes max étant commun aux deux méthodes dans leur variante Exhaustive.

3.2.5 Méthode Gain Ratio

Le gain ratio est la mesure utilisée dans la méthode C4.5 (Quinlan 1993). Le gain d'entropie suite à regroupement de modalités, utilisé dans l'algorithme ID3 précurseur de C4.5, est une mesure qui tend à favoriser les partitions à grand nombre de groupes. De fait, le gain d'entropie ne peut que décroître suite au regroupement de modalités, ce qui rend ce critère inutilisable pour le groupage. Le gain ratio apporte un correctif au gain d'entropie en le divisant par l'entropie des groupes. Si la réduction du gain est plus faible que la diminution de l'entropie des groupes, le gain ratio résultant peut augmenter, ce qui permet de rechercher des groupages pertinents. Il est à noter que (Elomaa & Rousu 1997) ont montré que le critère du Gain Ratio n'est pas "well-behaved" dans le cas des k-partitions, ce qui peut conduire à la séparation en différents groupes de valeurs descriptives mono-classe ayant même classe cible.

Une lecture attentive du chapitre consacré au groupage des attributs dans (Quinlan 1993) montre que Quinlan a rajouté une nouvelle contrainte à l'algorithme, en imposant que la partition finale ait un gain d'entropie supérieur à la moitié du gain d'entropie de la partition initiale. Un examen approfondi du code de C4.5, également publié dans (Quinlan 1993), montre qu'un prétraitement additionnel est effectué pour fusionner préalablement toutes les modalités initiales mono-classe ayant même classe cible (ce qui en effet ne serait pas garanti par l'optimisation du gain ratio qui n'est pas "well-behaved"). En définitive, nous avons réimplémenté l'algorithme glouton d'optimisation du gain ratio, en intégrant les spécificités mises en œuvre dans C4.5 (qui en pratique tendent à améliorer les résultats de l'heuristique de groupage), à savoir:

- Prétraitement: Regroupement des valeurs descriptives initiales mono-classe ayant même classe cible
- Mémorisation du gain originel: entropie de la classe cible – entropie de la classe cible après groupage
- Algorithme de groupage:
  - Tant que amélioration du gain ratio et que le gain résultant est supérieur à la moitié du gain originel, fusionner le couple de modalité apportant la meilleure amélioration

La méthode Gain Ratio ne peut produire que des partitions ayant au moins deux groupes, comme pour la méthode Tschuprow. Par ailleurs, on peut noter que la méthode Gain Ratio est la seule dont le critère d'évaluation d'une partition ne soit pas cumulatif, c'est-à-dire qu'il ne peut pas se décomposer sur l'ensemble des groupes de la partition. Cette particularité empêche la bufferisation des calculs intermédiaires qui pour les autres méthodes permet de se ramener à une complexité algorithmique en $O(I^2 \log(I))$. Dans le cas de la méthode Gain Ratio, cette complexité est en $O(I^3)$.

3.3 Jeux d'essai synthétiques

L'intérêt des jeux d'essai synthétiques est de contrôler exactement la distribution des modalités sources et des classes cibles, ce qui permet une évaluation optimale en test des groupages produits sur des échantillons d'apprentissage. Un jeu d'essai synthétique comportant un attribut descriptif et une classe cible est complètement défini par l'ensemble des paramètres suivants:

I: nombre de valeurs descriptives

J: nombre de classe cibles

$p_i, 1 \le i \le I$: distribution de probabilités des valeurs descriptives

$p_{j/i}, 1 \le j \le J,\ 1 \le i \le I$: distribution de probabilités des classes cibles conditionnellement aux valeurs descriptives

Dans cette section, on va étudier deux type de jeux d'essai synthétiques, dans le cas d'indépendance entre attribut descriptif et classe cible et dans un cas de dépendance contrôlée. Dans chacun des cas, on limitera l'étude au cas d'une équidistribution des modalités sources et des classes cibles, et aux cas avec 2 classes cibles et 4 classes cibles.

3.3.1 Groupage d'un attribut descriptif indépendant de l'attribut à prédire

Dans cette expérimentation, on utilise un attribut symbolique indépendant de l'attribut à prédire ($p_i = 1/I$ et $p_{j/i} = 1/J$). On utilise des échantillons d'apprentissage de taille 1000 dans les cas de 2 classes cibles et 4 classes cibles, en faisant varier le nombre de valeurs descriptives à grouper. L'expérimentation est menée 1000 fois sur des échantillons générés aléatoirement pour chaque type de jeux d'essai. Dans le cas d'indépendance, le nombre de groupes optimal est égal à 1. Pour chaque méthode évaluée, on comptabilise le nombre de groupes surnuméraires produits en moyenne. La qualité des partitions est évaluée en utilisant la divergence de Kullback-Leibler. Les jeux d'essai étant théoriques, les probabilités conditionnelles réelles des classes cibles sont connues de façon exacte, ce qui permet de calculer la divergence de Kullback-Leibler sans utiliser d'échantillon de test. Les probabilités conditionnelles estimées grâce au groupage sont évaluées par l'estimateur de Laplace sur l'ensemble d'apprentissage. La figure 2 présente le nombre moyen de groupes surnuméraires et la divergence de Kulback-Leibler des groupages en fonction de nombre de modalités sources, dans le cas de 2 classes cibles. La figure 3 présente les mêmes informations dans le cas de 4 classes cibles.

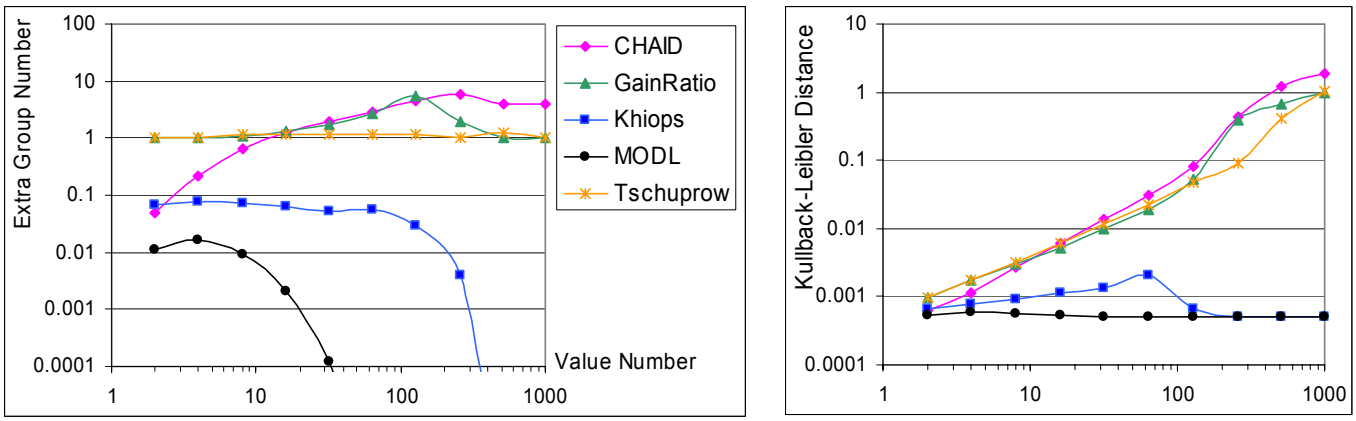

Figure 2 : Nombre moyen de groupes surnuméraires et divergence de Kulback-Leibler en fonction du nombre de modalités sources, dans le cas de deux classes cibles indépendantes des modalités sources, pour un échantillon de taille 1000

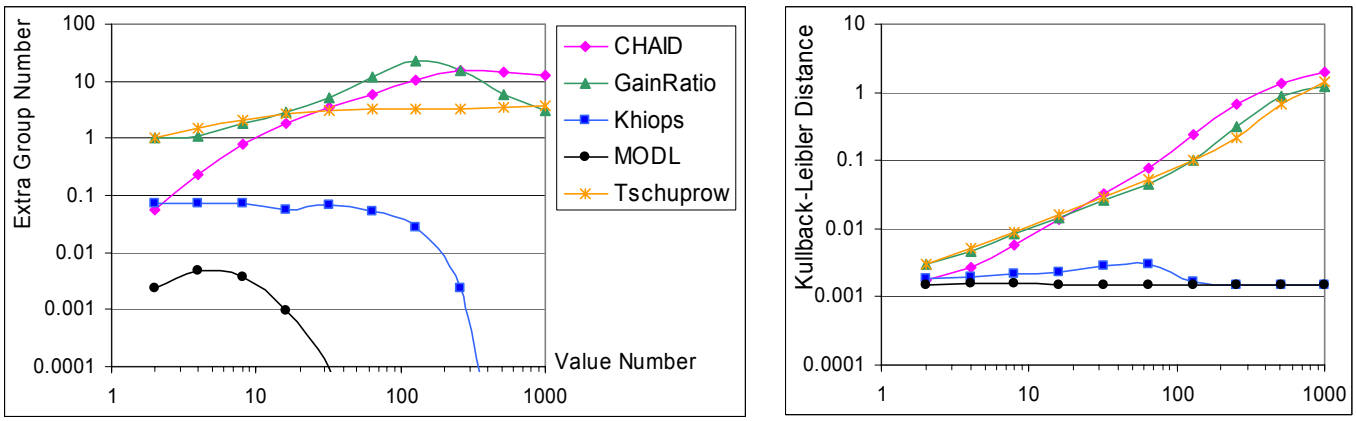

Figure 3 : Nombre moyen de groupes surnuméraires et divergence de Kulback-Leibler en fonction du nombre de modalités sources, dans le cas de quatre classes cibles indépendantes des modalités sources, pour un échantillon de taille 1000

Le premier enseignement de ces résultats est l'intérêt de la divergence de Kullback-Leibler pour estimer la qualité d'un groupage. Le taux de bonne prédiction est uniformément égal à 50% en test quelque soit le nombre de groupes produits et ne peut donc pas être utilisé comme critère d'évaluation. Dans le cas d'un attribut indépendant, le groupage optimum consiste à ne faire qu'un seul groupe, alors que le pire groupage consiste à ne rien grouper. Cela se traduit par une valeur de DKL d'autant meilleure que le nombre de groupes produits est faible. Cette amélioration provient du fait qu'après un groupage, l'effectif du groupe est accru, ce qui permet d'améliorer l'estimation de la probabilité conditionnelle des classes cibles.

La méthode CHAID produit d'autant plus de groupes que le nombre de modalités sources est important. Ce nombre de groupes croit également avec le nombre de classes cibles. Ce comportement de sur-apprentissage se traduit par une estimation DKL des probabilités des classes cibles se dégradant très rapidement avec l'augmentation du nombre de modalités sources.

La méthode GainRatio est contrainte à produire au moins deux groupes. Elle exhibe un comportement de sur-apprentissage en produisant de plus en plus de groupes dès qu'il y a plus d'une dizaine de modalités sources à grouper.

La méthode Tschuprow est également contrainte à produire au moins deux groupes. Elle limite le sur-apprentissage en produisant un nombre quasiment constant de groupes, égal au nombre de classes cibles. Ce comportement est conforme au biais du critère de Tschuprow qui ne peut atteindre son maximum que pour les tables de contingences carrées (ayant un nombre de groupe égal au nombre de classes cibles).

La méthode Khiops contrôle le sur-apprentissage de façon conforme à ses objectifs, en ne produisant qu'un seul groupe terminal dans environ 95% des cas, quelle que soit la nature du jeu d'essai. Les cas de partitions multi-groupes pour Khiops comportent systématiquement deux groupes. Quand le nombre de valeurs descriptives devient très important, la contrainte d'effectif minimum de Khiops devient active. Avant ce seuil (d'environ 100 valeurs descriptives), Khiops produit 1 seul groupe dans 95% des cas. Apres ce seuil, toutes les modalités sont inconditionnellement regroupées en un seul groupe.

La méthode MODL produit presque toujours des partitions mono-groupe, comme on s'y attend intuitivement d'après les propriété théoriques du critère utilisé. Pour MODL, les expérimentations ont été menées 100000 fois pour améliorer la qualité de l'estimation. Au delà de 50 valeurs descriptives, aucun groupage (sur 100000) n'a abouti a plus de un groupe.

En résumé, on observe ici trois types de comportements vis-à-vis du sur-apprentissage. La méthode MODL contrôle automatiquement de façon optimale son comportement et reconnaît presque toujours le cas d'indépendance. La méthode Khiops utilise un paramètre utilisateur lui permettant de contrôler le taux de sur-apprentissage. Toutes les autres méthodes sur-apprennent, d'autant plus que le nombre de modalités sources (également le nombre de classes cibles) est important.

3.3.2 Groupage d'un attribut descriptif dépendant de l'attribut à prédire

Dans cette expérimentation, on utilise une relation de dépendance contrôlée entre l'attribut descriptif et la classe à prédire. On utilise des groupes associés à une classe cible privilégiée majoritaire, les autres classes cibles étant équidistribuées. On introduit un paramètre M correspondant à un nombre de taux de mélange. Les taux de mélanges p_m sont calculés par $p_m = m/(M+1),\ 1 \le m \le M$. On utilise un nombre de groupes $G=M*J$ et des valeurs descriptives affectées en nombre égal à chaque groupe (I est un multiple de G). Dans chaque groupe g_{m_0,j_0}, la classe cible j_0 est privilégiée: chaque modalité source affectée au groupe est attribuée en priorité la classe cible privilégiée selon le taux de mélange ($p_{j_0/i} = p_{m_0} + (1-p_{m_0})/J$)et uniformément sur toutes les classes cibles sinon ($p_{j/i} = (1-p_{m_0})/J,\ \forall j \neq j_0$). La figure 4 illustre le cas de 4 classes cibles et 2 taux de mélanges, représentant un total de 8 groupes caractérisés par des distributions spécifiques des classes cibles.

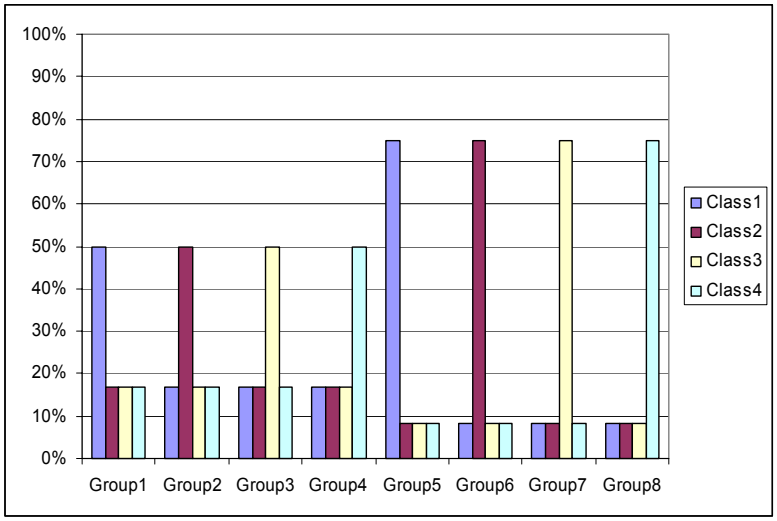

Figure 4 : Distribution des classes cibles dans chacun des 8 groupes d'un jeu d'essai synthétique caractérisé par 4 classe cibles et 2 taux de mélange. Les I valeurs descriptives sont uniformément distribuées sur les 8 groupes

On utilise des échantillons d'apprentissage de taille 10000 dans les cas de 2 classes cibles et 4 classes cibles avec 2 taux de mélanges (correspondant à 4 groupes et 8 groupes) et on fait varier le nombre de valeurs descriptives en prenant des multiples du nombre de groupes pour toujours maintenir une équidistribution des valeurs descriptives sur les groupes. L'expérimentation est menée 1000 fois sur des échantillons générés aléatoirement pour chaque type de jeux d'essai.

Pour chaque méthode évaluée, on comptabilise le nombre moyen de groupes produits. La qualité des partitions est évaluée en utilisant la divergence de Kullback-Leibler. La figure 5 présente les résultats dans le cas de 2 classes cibles pour 4 groupes théoriques et la figure 6 dans le cas de 4 classes cibles pour 8 groupes théoriques (ceux de la figure 4).

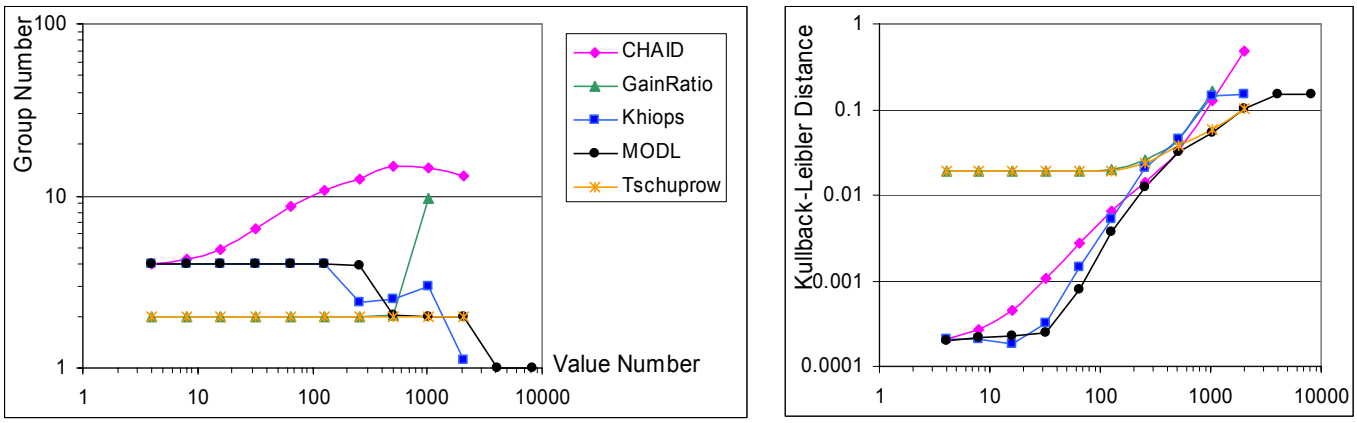

Figure 5 : Nombre moyen de groupes et divergence de Kulback-Leibler en fonction du nombre de modalités sources, dans le cas de deux classes cibles et 4 groupes synthétiques pour un échantillon de taille 10000

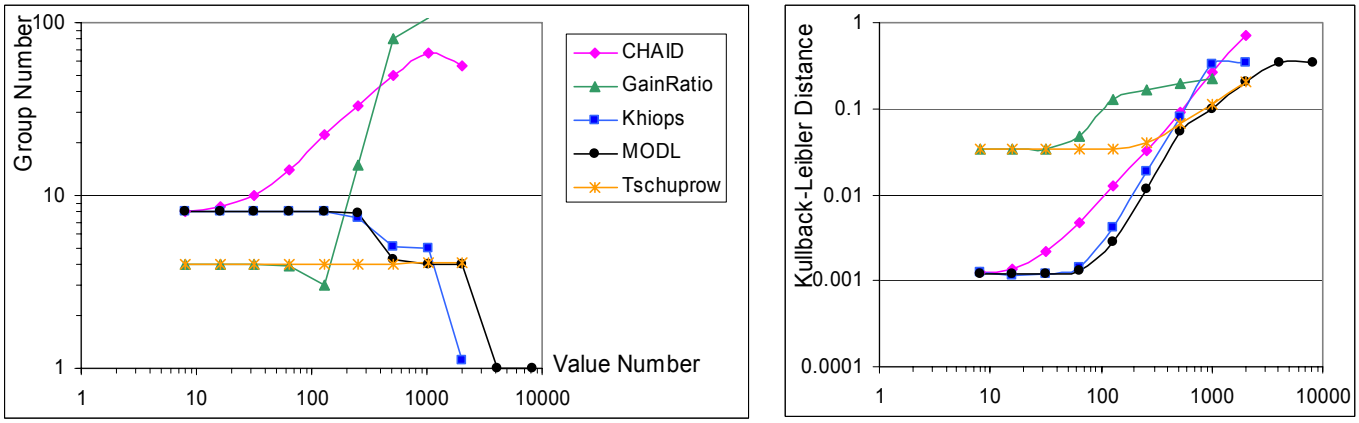

Figure 6 : Nombre moyen de groupes et divergence de Kulback-Leibler en fonction du nombre de modalités sources, dans le cas de quatre classes cibles et 8 groupes synthétiques pour un échantillon de taille 10000

Cette deuxième expérimentation confirme l'intérêt de la divergence de Kullback-Leibler comme mesure de qualité d'un groupage. On vérifie que sa valeur est parfaitement corrélée avec la qualité des groupages, dont l'optimum est ici connu de façon exacte.

Comme dans le cas d'indépendance, la méthode CHAID produit d'autant plus de groupes que le nombre de modalités sources est important. Ce nombre de groupe croit également avec le nombre de classes cibles.

La méthode GainRatio a un comportement surprenant. Alors que dans le cas d'indépendance, elle produit un nombre croissant de groupes quand le nombre de valeurs descriptives augmente, elle semble ici fortement biaisée en faveur d'un nombre de groupes égale au nombre de classes cibles. A partir d'une centaine de modalité, elle quitte ce comportement (de sous-apprentissage) pour un comportement de sur-apprentissage conduisant à un nombre croissant de groupes. Les expérimentations n'ont pas pu être menées au delà de 1000 modalités sources en raison de la complexité algorithmique de la méthode (cubique par rapport au nombre de modalités).

La méthode Tschuprow reste conforme à son biais en produisant un nombre quasiment constant de groupes égal au nombre de classes cibles. Cela se traduit ici par une mauvaise estimation des probabilités cibles.

La méthode Khiops bénéficie de son contrôle du sur-apprentissage et trouve automatiquement le nombre théorique de groupes quand les effectifs sont suffisants. A partir d'environ 300 modalités sources, la contrainte d'effectif minimum de Khiops devient active et le nombre de groupes produits chute rapidement à 1 quand la plupart des modalités sources n'atteignent pas un effectif suffisant pour assurer la fiabilité du test du Khi2.

La méthode MODL produit presque toujours des partitions optimales comportant le nombre théorique de groupes. Quand le nombre de modalités devient très important (environ 500, soit un effectif moyen de 40 par modalités), il se produit une transition et MODL ne produit plus qu'un groupe par classe cible. Autrement dit, l'effectif par modalité source ne permet plus de discerner tous les groupes théoriques et la méthode dégrade son comportement en diminuant le nombre de groupes séparables. Quand le nombre de modalités augmente encore (au delà de 2000 modalités, soit moins de 5 instances par modalité en moyenne), MODL ne produit qu'un seul groupe. MODL est la seule méthode pour laquelle les groupages ont été menés jusqu'à 10000 modalités (en raison de sa faible complexité algorithmique). En ce qui concerne l'évaluation de la distribution de

probabilités de classes cibles, on constate que la divergence de Kullback-Leibler pour MODL est systématiquement meilleure que la borne inf des autres méthodes évaluées.

Afin d'évaluer l'impact d'un déséquilibre dans la distribution des effectifs des modalités sources, on a procédé à une variante de l'expérimentation avec 4 classes cibles et 8 groupes synthétiques. On s'est focalisé sur des jeux d'essai comportant exactement 128 modalités sources dispatchées sur les 8 groupes synthétiques. Les effectifs des modalités sources sont en progression géométrique, la modalité la plus fréquente (environ 3,5% des instances) étant 100 fois plus fréquente que la modalité la plus rare. On a fait varier la taille de l'échantillon de 100 instances (pratiquement aucune modalité n'a un effectif significatif) à 10000 (pratiquement toutes les modalités ont un effectif significatif). L'expérimentation permet de mesurer l'apport de la gestion du groupe poubelle, indiscernable dans les cas de modalités équidistribuées. Il est à noter que la progression géométrique représente un cas difficile pour ajuster l'effectif des modalités hors poubelle: les cas avec une transition nette entre les modalités fréquentes et les modalités rares sont naturellement plus adaptés à la poubellisation des modalités rares. La figure 7 présente les résultats de cette expérimentation.

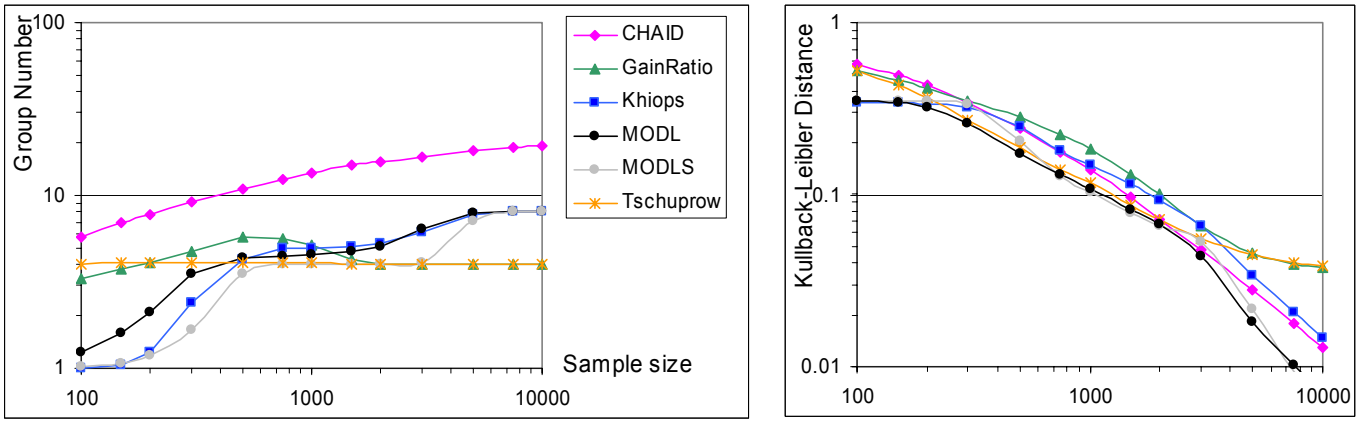

Figure 7 : Nombre moyen de groupes et divergence de Kulback-Leibler en fonction de la taille de l'échantillon, dans le cas de 128 modalités sources de distribution fortement déséquilibrée, quatre classes cibles et 8 groupes synthétiques

On retrouve un comportement similaire à celui des expérimentations précédentes pour l'ensemble des méthodes évaluées. La méthode MODLStandard ne produit d'abord qu'un seul groupe quand la taille de l'échantillon est petite devant le nombre de modalités à traiter. A partir d'environ 500 instances, la méthode produit 4 groupes (un par classe cible). Au delà de 5000 instances, les 8 groupes synthétiques sont correctement identifiés. La méthode MODL bénéficie de son groupe poubelle en procédant aux mêmes transitions dans les nombres de groupes que MODLStandard, mais systématiquement en ayant besoin de moins d'instances (de 50% à 100% en moins). La prise en compte de la poubelle résulte d'un compromis: la poubelle tend d'une part à diminuer la qualité prédictive en ayant un groupe mélangeant tout type de modalités (comportement non "well-behaved"), d'autre part à améliorer la qualité en permettant plus tôt des groupages efficaces. L'évaluation globale de la qualité par la divergence de Kullback-Leibler montre l'apport de la prise en compte de la poubelle. La méthode MODL est systématiquement plus performante que la meilleure des autres méthodes, ce qui n'est pas le cas pour la méthode MODLStandard.

En conclusion, l'ensemble des jeux d'essai d'indépendance et de dépendance contrôlée a permis d'exhiber une grande variété de comportement des méthodes de groupage testées. La méthode CHAID sur-apprend en produisant systématiquement trop de groupes. Les méthodes GainRatio et Tschuprow sont fortement biaisées par leur critère d'évaluation des groupages et tendent à produire régulièrement un nombre de groupes égal au nombre de classes cibles. La méthode Khiops contrôle le sur-apprentissage au moyen d'un paramètre utilisateur, mais sa contrainte d'effectif minimum liée à une fiabilisation du test du Khi2 limite son domaine de validité au cas où les modalités sources ne sont pas trop nombreuses. La méthode MODL obtient des résultats apparemment optimaux et conformes à ses fondements théoriques, en produisant le groupage le plus probable connaissant les données.

3.3.3 Performance CPU

L'expérimentation consiste à évaluer le temps CPU de groupage. La complexité algorithmique de chaque méthode a déjà été étudiée: $O(n.\log(n))$ pour MODL, $O(n^3)$ pour GainRatio et $O(n^2.\log(n))$ pour autres méthodes (CHAID, Tschuprow et Khiops). Il est néanmoins intéressant de mesurer le temps CPU de groupage en pratique, notamment pour MODL qui incorpore plusieurs étapes de pré-optimisation et post-optimisation. On ne s'intéresse pas à la performance absolue (dépendante de la qualité de l'implémentation et de la puissance de la machine), mais plutôt à la variation de la performance par rapport aux nombres de valeurs descriptives.

On utilise ici des jeux d'essai comportant 4 classes cibles et 8 groupes synthétiques. On fait varier le nombre de modalités sources, la taille des échantillons étant égale à 100 fois le nombre de modalités sources. Les résultats sont présentés sur la

figure 8, en reportant le temps de tri des instances à titre de comparaison.

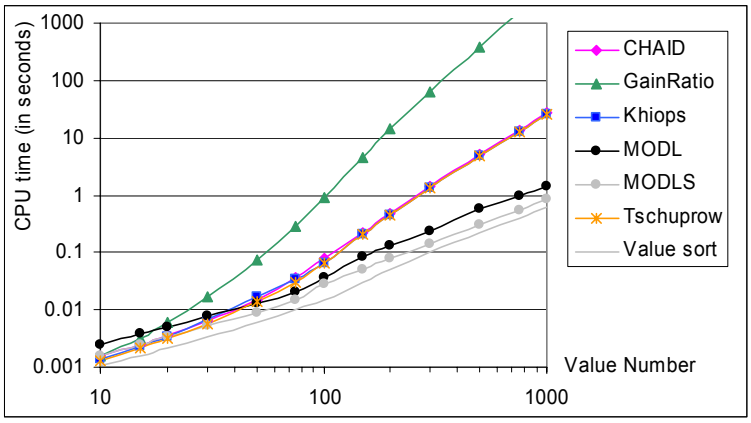

Figure 8 : Temps de groupage (en secondes) en fonction du nombre de modalités sources, dans le cas d'un échantillon ayant 100 instances par modalités sources, 4 classes cibles et 8 groupes synthétiques

Les résultats se catégorisent très nettement en trois familles, correspondant aux comportements théoriques super-linéaire, quadratique et cubique vis-à-vis du nombre de modalités sources. Au delà de quelques dizaines de modalités sources, la méthode MODL domine les autres méthodes en performance CPU en dépit du surcoût dû à ses étapes de pré et post-optimisation. Dès 100 modalité sources, la méthode MODL est environ 2 fois plus rapide que la méthode CHAID, 4 fois plus rapide pour 200 modalités. Le temps CPU nécessité par MODL reste toujours du même ordre de grandeur que le temps de tri des instances par valeur descriptive. Plus précisément, la méthode MODLStandard est environ 50% plus lente que le temps de tri des instances (ici 100 instances par modalité); la méthode MODL est elle-même environ 50% plus lente que la méthode MODLStandard, quel que soit le nombre de modalités.

3.4 Benchmarks UCI

3.4.1 Méthodologie

Les méthodes de groupage présentées sont évaluées en utilisant douze jeux de données standards extraits de la base UCI (Blake and Merz 1998), comportant au moins quelques dizaines d'instances par classe cible et des attributs symboliques non réduits à deux modalités. Afin d'augmenter le nombre d'attributs à grouper, les attributs continus ont également été soumis au groupage après une discrétisation non supervisée préalable en 10 intervalles de largeur égale appliquée au jeu de données initial complet. Les 12 jeux de données, totalisant 230 attributs à grouper, sont résumés dans le tableau 2, dont la dernière colonne présente le taux de bonne prédiction du prédicteur majoritaire.

Tableau 2 : Jeux de données de l'UCI utilisés pour les expérimentations de groupage

Dataset	Continuous Attributes	Nominal Attributes	Size	Class Values	Majority Accuracy
Adult	7	8	48842	2	76,07
Australian	6	8	690	2	55,51
Breast	10	0	699	2	65,52
Crx	6	9	690	2	55,51
Heart	10	3	270	2	55,56
HorseColic	7	20	368	2	63,04
Ionosphere	34	0	351	2	64,10
Mushroom	0	22	8416	2	53,33
TicTacToe	0	9	958	2	65,34
Vehicle	18	0	846	4	25,77
Waveform	40	0	5000	3	33,84
Wine	13	0	178	3	39,89

Les attributs symboliques des jeux de données comportent en moyenne moins d'une dizaine de valeurs descriptives, ce qui s'avère peu sélectif (cf. jeux d'essai synthétiques). On a constitué une seconde base de jeux de données bivariés en générant tous les produits cartésiens des attributs descriptifs initiaux. Cette seconde base comporte 2614 attributs ayant en moyenne 55 valeurs descriptives.

L'expérimentation consiste à grouper chaque attribut puis à mesurer le nombre de groupes résultant et le taux de bonne prédiction en test. Le taux de bonne prédiction en apprentissage est également mesuré afin d'évaluer la robustesse de chaque méthode de groupage par l'écart entre apprentissage et test. La divergence de Kullback-Leibler est enfin utilisée pour évaluer la qualité de l'estimation de la distribution de probabilité des classes cibles. L'impact sur la classification est étudié au moyen du prédicteur Bayesien Naïf. Les mesures sont effectuées en utilisant la procédure de validation croisée stratifiée en 10 étapes. Toutes les méthodes testées ont été réimplémentées, afin d'effectuer les tests sans biais potentiel dû au choix des coupures dans la validation croisée. Les critères sont mesurés par leur moyenne globale dans le cas des jeux d'essai univariés (2300 points de mesure) et bivariés (26140 points de mesures). En l'absence de connaissance des distributions théoriques des jeux d'essais, l'évaluation des méthodes par leur comportement moyen permet une comparaison macroscopique des méthodes.

Les résultats des expérimentations sont étudiés au moyen de l'analyse multicritère, dont nous rappelons tout d'abord quelques notions. On dit qu'un solution *domine* (ou est *non inférieure* à) une autre solution si elle est meilleure sur tous les critères. Une solution ne peut être dominée si toute amélioration sur un des critères entraîne une détérioration sur un autre critère. Une telle solution est un *optimum de Pareto*. La *surface de Pareto* (*courbe de Pareto* pour deux critères) est l'ensemble de tous les optima de Pareto. Les figures 9 à 12 présentent les résultats agrégés de l'expérimentation sur des plans bi-critères. Les résultats en univarié sont représentés sur les graphiques de gauche et en bivarié sur ceux de droite. Sur chaque graphique, la courbe de Pareto est visualisée en grisé afin de situer rapidement les méthodes.

3.4.2 Nombre de groupes et taux de bonne prédiction

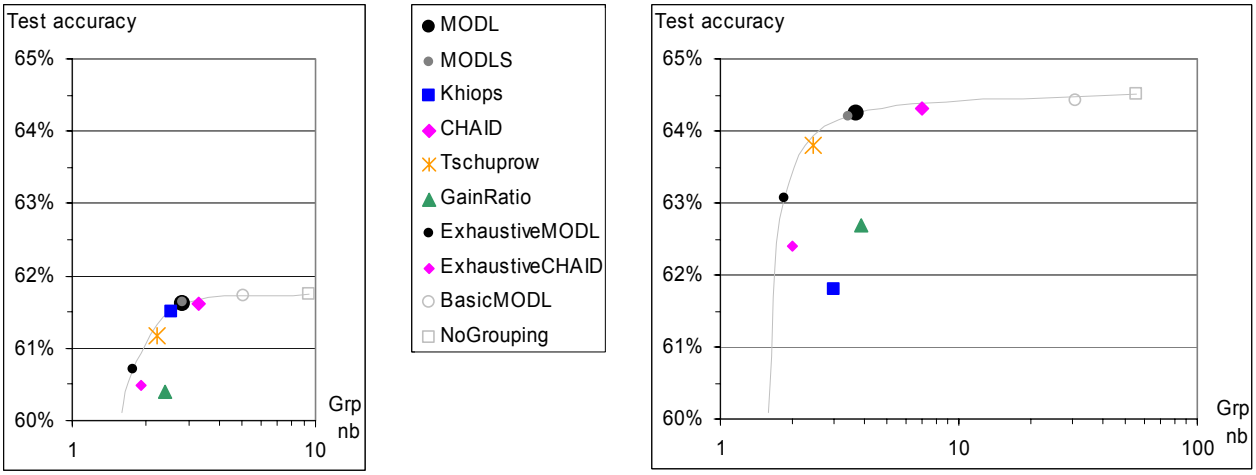

Figure 9 : Nombre moyen de groupes et taux de bonne prédiction en test des partitions produites par les méthodes de groupage testées sur les bases de l'UCI, en univarié (à gauche) et en bivarié (à droite)

La figure 9 analyse le lien entre le nombre de groupes produits et le taux de bonne prédiction en test. En terme de performance prédictive uniquement, la méthode consistant à ne rien grouper (NoGrouping) est la meilleure, au détriment d'un nombre de groupes maximal. En analyse univariée, la plupart des méthodes se retrouvent sur la courbe de Pareto et représentent différents compromis optimaux (parmi les méthodes testées) entre nombre de groupes et taux de bonne prédiction en test. Seule GainRatio est largement dominée par d'autres méthodes, particulièrement par la méthode ExhaustiveMODL pourtant limitée à deux groupes. La méthode ExhaustiveMODL domine également son homologue ExhaustiveChaid. La méthode Tschuprow semble Pareto optimale. Son fort biais en faveur de groupages ayant un nombre de groupes égal au nombre de classe cibles ne la pénalise que faiblement en taux de bonne prédiction. La méthode BasicMODL arrive à réduire de façon importante le nombre de modalités avec une perte insignifiante en qualité prédictive. En analyse univariée, les méthodes Khiops, MODL et CHAID sont très proches. En analyse bivariée, les contrastes de performances entre méthodes sont accentués du fait de la complexité accrue des groupages à effectuer. La performance de la méthode Khiops chute du fait de sa contrainte d'effectif minimum qui l'oblige à mélanger les modalités rares pour assurer la fiabilité du test du Khi2. La méthode CHAID est à peine plus performante que la méthode MODL, au prix d'un nombre de groupes moyen environ deux fois plus élevé. La méthode MODL se situe à une position intéressante de la courbe de Pareto. La pente de la courbe est forte avant MODL est faible après. Ainsi, une faible diminution du nombre de groupe entraîne une perte significative en qualité prédictive, alors qu'une forte augmentation du nombre de groupe n'apporte que peu d'amélioration.

3.4.3 Nombre de groupes et qualité de l'estimation des distributions cibles

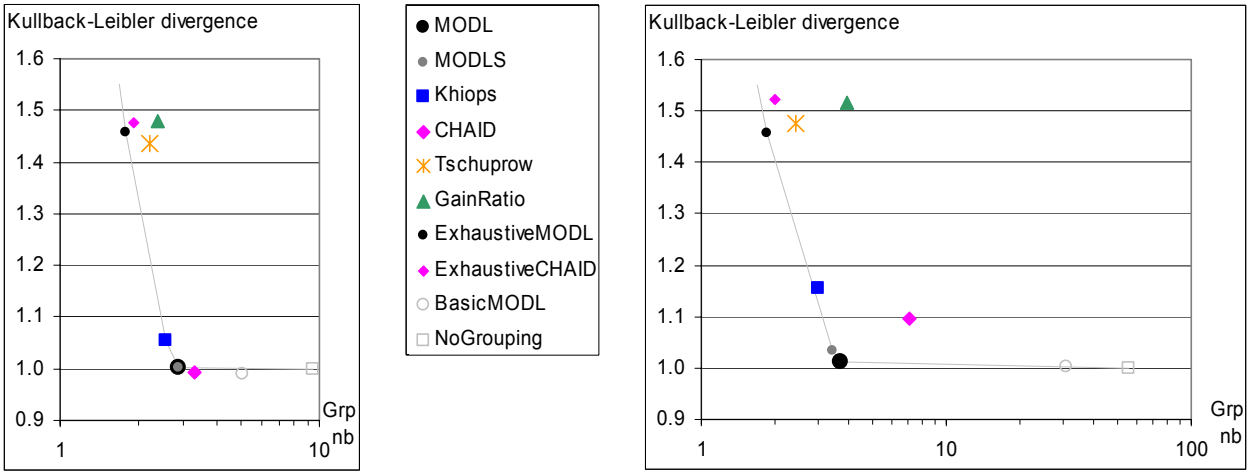

Figure 10 : Nombre moyen de groupes et qualité prédictive (divergence de Kullback-Leibler) des partitions produites par les méthodes de groupage testées sur les bases de l'UCI, en univarié (à gauche) et en bivarié (à droite)

La figure 10 se focalise sur le lien entre le nombre de groupes et la qualité de l'estimation des probabilités cibles. La divergence de Kullback-Leibler étant une mesure de la qualité prédictive plus appropriée que le taux de bonne prédiction en test, on retrouve des résultats similaires à ceux de la figure 9 en les affinant. La encore, ne rien grouper est optimal si l'on ne s'intéresse qu'à la qualité prédictive. Néanmoins, en univarié, les méthodes CHAID, MODL et Khiops sont pratiquement optimales en qualité prédictive avec beaucoup moins de groupes. La méthode Tschuprow est cette fois pénalisée par son biais sur le nombre de groupes qui diminue sa performance au niveau des méthodes contraintes à au plus deux groupes (ExhaustiveCHAID et Exhaustive MODL). En bivarié, les méthodes se différencient davantage. La méthode Khiops qui sous-apprend en raison de sa contrainte d'effectif minimum assurant la fiabilité du Khi2 décroche nettement de l'évaluation prédictive optimale. De même, la méthode CHAID qui elle sur-apprend en produisant trop de groupes est pénalisée par une diminution de sa performance prédictive. Seule la méthode MODL trouve le compromis optimal (parmi les méthodes testées, en moyenne sur les jeux de données de l'expérimentation).

3.4.4 Taux de bonne prédiction en apprentissage et en test

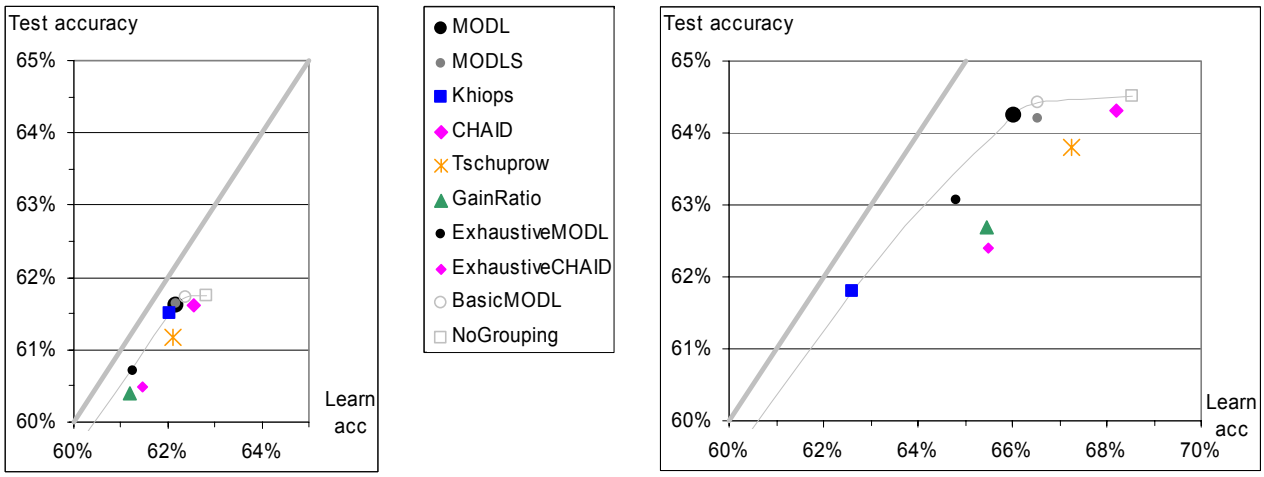

Figure 11 : Taux de bonne prédiction en apprentissage et en test pour les partitions produites par les méthodes de groupage testées sur les bases de l'UCI, en univarié (à gauche) et en bivarié (à droite)

La figure 11 représente la robustesse des méthodes en reportant conjointement le taux de bonne prédiction en test et en apprentissage. Ici, la diagonale (en grisé épais) représente la robustesse optimale théorique (approchée avec des groupages mono-groupe, au prix d'une performance prédictive réduite à celle du prédicteur majoritaire). Le compromis est ici entre robustesse (faible distance à la diagonale) et taux de bonne prédiction en test. Les résultats montrent que la méthode MODL (et toutes ses sous-variantes) sont robustes. Khiops l'est également aux prix d'un sous-apprentissage. Toutes les autres

méthodes ont une robustesse comparable, nettement inférieure à celle de MODL. On remarque également que la méthode BasicMODL apporte une amélioration significative par rapport à la méthode ne groupant rien. On note enfin un léger avantage en robustesse de MODL par rapport à sa variante sans poubelle MODLStandard.

3.4.5 Impact sur le prédicteur Bayesien Naïf

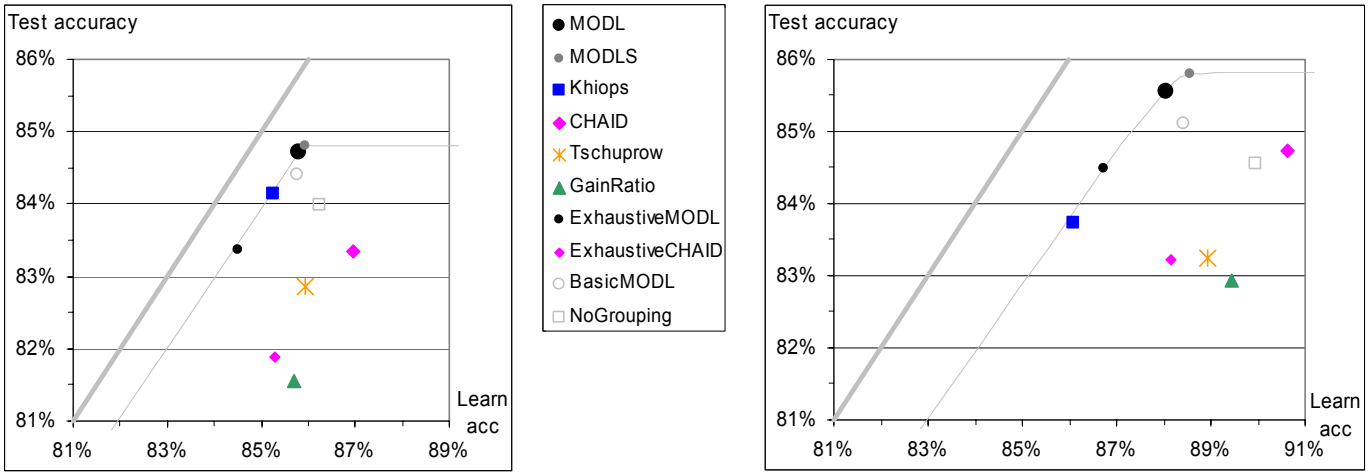

Figure 12 : Nombre moyen de groupes et performance prédictive des partitions produites par les méthodes de groupage testées sur les bases de l'UCI, en univarié (à gauche) et en bivarié (à droite)

Afin d'évaluer l'impact des méthodes de groupage sur la classification, nous les avons utilisées pour prétraiter les attributs symboliques dans un prédicteur Bayesien Naïf. Le prédicteur Bayesien Naïf (Langley et al 1992) prédit pour chaque instance la classe cible la plus probable conditionnellement aux attributs descriptifs, en faisant l'hypothèse d'indépendance entre les attributs descriptifs pour chaque classe cible. Après l'étape de groupage, les probabilités des attributs symboliques sont estimées par comptage (en utilisant l'estimateur de Laplace). L'évaluation a été menée sur les douze jeux de données de l'UCI utilisés dans l'expérimentation, d'une part en univarié (avec les attributs initiaux des bases), d'autre part en univarié (en prenant toutes les paires d'attributs). Les résultats sont présentés sur la figure 12 avec la performance en apprentissage et en test afin d'évaluer la robustesse des méthodes. La méthode MODL bénéficie de la qualité de ses groupages pour dominer toutes les autres méthodes. Le prédicteur Bayesien Naïf résultant est de loin le plus robuste tout en améliorant la performance prédictive des autres méthodes, notamment de la méthode consistant à estimer directement les probabilités à partir des valeurs descriptives (sans aucun groupage). Le groupage a été principalement étudié et appliqué dans le cadre des arbres de décision afin de limiter une diminution trop rapide du nombre d'instances dans les sous-branches de l'arbre. L'expérimentation montre également l'intérêt du groupage en tant que méthode de prétraitement pour le prédicteur Bayesien. Par la même occasion, on constate une amélioration globale des performances pour l'ensemble des méthodes quand on passe de l'univarié au bivarié. Le bivarié permet d'envisager certaines interactions simples entre attributs descriptifs et d'aller ainsi au delà de l'hypothèse limitative d'indépendance entre attributs du prédicteur Bayesien Naïf. Cette amélioration est plus importante si les valeurs descriptives sont groupées (gain moyen d'environ 1%) que si elles ne le sont pas (gain de 0.5%).

Conclusion

La méthode MODL repose sur une approche bayésienne du groupage. Un modèle standard de groupage est proposé, suivi d'une distribution a priori de ces modèles. On a défini un a priori à trois étages, pour lequel on choisit de façon uniforme d'abord le nombre de groupes, puis la partition des valeurs descriptives en groupes, enfin la distribution des classes cibles sur chaque groupe. On a en plus supposé l'indépendance des distributions entre les groupes. Sur la base de ces hypothèses, on a montré qu'il existe un critère d'évaluation permettant de rechercher le groupage optimal au sens de Bayes. Dans le cas à deux classes cibles, un algorithme optimal en $O(n^3)$ permet de trouver l'optimum global de ce critère et donc d'aboutir au groupage optimal. Dans le cas général, une heuristique en $O(n.\log(n))$, basée sur une méthode gloutonne ascendante et intégrant des étapes de pré-optimisation et post-optimisation est utilisable. On a également introduit la possibilité de ranger les modalités rares dans un groupe poubelle afin de se concentrer sur les modalités d'effectif significatif, et proposé pour cette extension un nouveau critère d'optimalité des groupages ainsi que des heuristiques d'optimisation performantes.

Des expérimentations comparatives intensives ont été menées sur de nombreux jeux de données de l'UCI ainsi que sur des jeux de données synthétiques. Les résultats ont montré que la méthode MODL domine les méthodes alternatives testées sur l'ensemble des critères évalués, à savoir le taux de bonne prédiction, la qualité de l'estimation de la distribution des classes cibles, la robustesse, le faible nombre de groupes produits, la résistance au bruit, le seuil de détection des groupes pertinents, l'absence de paramétrage, le domaine d'applicabilité et la performance CPU de groupage.

Références

Berckman, N.C. (1995). Value grouping for binary decision trees. Technical Report, Computer Science Department – University of Massachusetts.

Blake, C.L. et Merz, C.J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

Boullé, M. (2003). Groupage robuste des valeurs d'un attribut symbolique par la méthode Khiops. Note technique NT/FTR&D/8028; France Telecom R&D.

Boullé, M. (2004). MODL: une méthode quasi-optimale de discrétisation supervisée. Note technique NT/FTR&D/8444; France Telecom R&D.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). Classification and Regression Trees. California : Wadsworth International.

Cestnik, B., Kononenko, I. & Bratko, I. (1987); ASSISTANT 86: A knowledge-elicitation tool for sophisticated users. In Bratko & Lavrac (Eds.), *Progress in Machine Learning*. Wilmslow, UK: Sigma Press.

Chou, P.A. (1991). Optimal Partitioning for Classification and Regression Trees. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 13, No. 4, p. 340-354.

Elomaa, T. & Rousu, J. (1997). Well-Behaved Evaluation Functions for Numerical Attributes. ISMIS'97, p. 147-156.

Fulton, T., Kasif, S., and Salzberg, S. (1995). Efficient algorithms for finding multi-way splits for decision trees. In Proc. Thirteenth International Joint Conference on Artificial Intelligence, p. 244-255. San Francisco, CA: Morgan Kaufmann.

Kass G.V. (1980). An exploratory technique for investigating large quantities of categorical data. Applied Statistics, 29(2) : 119-127.

Kerber R. (1991). Chimerge discretization of numeric attributes. Proceedings of the 10th International Conference on Artificial Intelligence, p. 123-128.

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of bayesian classifiers. *Proceedings of the 10th national conference on Artificial Intelligence*, AAAI Press, 223-228.

Lechevallier, Y. (1990). Recherche d'une partition optimale sous contrainte d'ordre total. Technical report N°1247. INRIA.

Quinlan J.R. (1986). Induction of decision trees. Machine Learning, 1, p. 81-106.

Quinlan J.R. (1993). C4.5 : Programs for Machine Learning. Morgan Kaufmann.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. Ann. Statist. 11, p 416-431.

Ritschard, G., Zighed, D.A. & Nicoloyannis, N. (2001). Maximisation de l'association par regroupement de lignes ou de colonnes d'un tableau croisé. Math. & Sci. Hum., n°154-155, p. 81-98.

Zighed, D.A. et Rakotomalala, R. (2000), Graphes d'induction. HERMES Science Publications, p. 327-359.

4 Annexe

4.1 A priori universel sur les entiers

Dans ce paragraphe, on décrit et commente l'a priori universel sur les entiers présenté par (Rissanen 1983).

Quand un entier appartient à un ensemble de taille finie *N*, on peut utiliser un a priori uniforme, où tout entier a la même probabilité de tirage 1/*N*. Cette approche ne peut s'appliquer dans le cas des ensembles infinis, pour lesquels Rissanen a proposé un a priori universel. Cet a priori universel pour les entiers est défini de telle façon que les petits entiers soient plus probables que les grands, et que le taux de décroissance des probabilités de tirage soit contraint à être le plus faible possible. Selon Rissanen, cet a priori est "universel" dans le sens où les longueurs de codage correspondantes (opposé du log des probabilités) correspondent au codage le plus compact des grands entiers. Cet a priori parait intéressant même dans le cas d'ensembles finis d'entiers, parce qu'il rend les petits entiers préférables aux grands avec la plus légère différence possible.

La longueur de code de l'a priori universel est donnée par

$$L(n) = \log_2(c_0) + \log_2^*(n) = \log_2(c_0) + \sum_{j>1} \max\left(\log_2^{(j)}(n), 0\right)$$

où $\log_2^{(j)}(n)$ est la j[th] composition de $\log_2$ ($\log_2^{(1)}(n) = \log_2(n)$, $\log_2^{(2)}(n) = \log_2(\log_2(n))$ …) et $c_0 = \sum_{n>1} 2^{-\log_2^*(n)} = 2.865...$

L'a priori universel sur les entiers est alors $p(n) = 2^{-L(n)}$.

4.2 Critère MODL standard d'évaluation des groupage

Notations:
n: nombre d'instance de la chaîne à grouper,
J: nombre de classes,
I: nombre de valeurs descriptives,
n_i: nombre d'instances pour la valeur descriptive *i,*
n_{ij}: nombre d'instances de la classe *j* pour la valeur descriptive *i*,
K: nombre de groupes,
k(i): index du groupe auquel est rattaché la valeur descriptive *i*,
n_k: nombre d'instances pour le groupe *k*,
n_{kj}: nombre d'instances de classe *j* pour le groupe *k*.

4.2.1 Préliminaires

On va énoncer ci-dessous quelques résultats de combinatoire, qui seront utilisés par la suite.

Lemme 1: Soit une urne contenant *n* boules indiscernables, et *k* urnes vides. Le nombre de partitions des *n* boules dans les *k* urnes est égal à C_{n+k-1}^{k-1}.

Preuve:
Soit n_1, n_2, … n_k le nombre de boules dans la $k^{\text{ième}}$ urne.
On s'intéresse ici uniquement au nombre de solutions distinctes de $n_1+n_2+ \ldots +n_k = n$.
L'expression précédente montre que ce problème revient au choix de (*k*-1) signes '+' dans une urne contenant *n* boules et *k*-1 signes '+'. Le nombre de répartitions recherché est donc C_{n+k-1}^{k-1}.

Coefficient multinomial:
On rappelle également un résultat sur la loi multinomiale. Le coefficient multinomial $n!/n_1!n_2!...n_k!$ correspond au nombre de répartitions distinguables de *n* boules dans *k* urnes contenant respectivement n_1, n_2, … n_k boules.

Partitions d'un ensemble:
Soit $S(n,k)$ le nombre de partition en *k* parties non vides d'un ensemble de *n* éléments. $S(n,k)$ est appelé le nombre de Stirling de deuxième espèce.

Soit $B(n,k)$ le nombre de partition en *k* parties (éventuellement vides) d'un ensemble de *n* éléments.
Pour *k*=*n*, ce nombre est égal au nombre total de partitions, ce qui correspond au nombre de Bell (noté $B(n)$).
Par la suite, $B(n,k)$ sera appelé nombre de Bell généralisé.

Par définition, on a $B(n,k)=\sum_{i=1}^{k} S(n,i)$.

4.2.2 Groupage avec a priori a trois étages

Théorème: Un modèle de groupage standard *M* suivant un a priori à trois étages est un modèle de prédiction optimal au sens de Bayes si son évaluation par la formule suivante est minimale sur l'ensemble de tous les modèles:

$$Value(M)=\log(I)+\log(B(I,K))+\sum_{k=1}^{K}\log\left(C_{n_k+J-1}^{J-1}\right)+\sum_{k=1}^{K}\log\left(n_k!/n_{k,1}!n_{k,2}!..n_{k,J}!\right)$$

Preuve:

Pour un groupage de type SGM, un a priori est défini dès que l'on connaît une distribution de probabilité de ses paramètres caractéristiques. Par soucis de simplification, on adoptera les notations suivantes:

p(*K*): probabilité a priori d'observer une valeur de *K*

p($\{k(i)\}$): probabilité a priori d'observer une partition (définie par $\{k(i)\}$) des valeurs descriptives en *K* groupes

p($\{n_{kj}\}$): probabilité a priori d'observer l'ensemble des valeurs n_{kj} pour un nombre de groupes *K* donné

p($\{n_{kj}\}_k$): probabilité a priori d'observer l'ensemble des valeurs n_{kj} d'un groupe donné *k*

L'objectif est de trouver le modèle de discrétisation *M* qui maximise la probabilité $p(M/S)$ pour une chaîne *S* données de classes cibles. En utilisant la formule de Bayes et en remarquant que la probabilité $p(S)$ est constante quelque soit le modèle, cela revient à maximiser $p(M)p(S/M)$.

Dans un premier temps, on va calculer la probabilité a priori $p(M)$ du modèle. On a

$$\begin{aligned} p(M) &= p\left(K,\{k(i)\},\{n_{kj}\}\right) \\ &= p(K)p\left(\{k(i)\}/K\right)p\left(\{n_{kj}\}/K,\{k(i)\}\right) \end{aligned}$$

Le nombre de groupes étant compris entre 1 et *I* de façon équiprobable, on a

$$p(K)=\frac{1}{I}.$$

Pour un nombre de groupes donné, toutes les partitions des *I* valeurs descriptives en *K* groupes (non nécessairement vides) sont équiprobables. On a

$$p\left(\{k(i)\}/K\right)=\frac{1}{B(I,K)}.$$

Le dernier terme à évaluer peut être réécrit comme un produit en utilisant l'hypothèse d'indépendance des distributions de classes cibles entre les groupes. On a

$$\begin{aligned} p\left(\{n_{kj}\}/K,\{k(i)\}\right) &= p\left(\{n_{kj}\}_1,\{n_{kj}\}_2,...,\{n_{kj}\}_K/K,\{k(i)\}\right) \\ &= \prod_{k=1}^{K} p\left(\{n_{kj}\}_k/K,\{k(i)\}\right) \\ &= \prod_{k=1}^{K} p\left(\{n_{kj}\}_k/n_k\right) \end{aligned}$$

(Les fréquences n_i par valeur descriptives étant connues, la connaissance de la partition des *I* valeurs descriptives en *K* groupes permet de déterminer l'effectif de chaque groupe.)

Pour un groupe *k* donné de taille n_k, toutes les distributions de classes cibles sont équiprobables. Le calcul de la probabilité d'une distribution est à nouveau un problème combinatoire dont la solution est:

$$p\left(\{n_{kj}\}_k/n_k\right)=\frac{1}{C_{n_i+J-1}^{J-1}}.$$

Donc,

$$p\left(\{n_{kj}\}/K,\{k(i)\}\right)=\prod_{k=1}^{K}1/C_{n_k+J-1}^{J-1}.$$

La probabilité a priori d'un modèle est alors

$$p(M)=\frac{1}{I}\frac{1}{B(I,K)}\prod_{k=1}^{K}\frac{1}{C_{n_k+J-1}^{J-1}}.$$

Evaluons maintenant la probabilité d'observer une chaîne *S* pour un modèle *M* donné. La probabilité de la chaîne est indépendante de l'ordre des instances. On peut donc trier les instances selon les groupes correspondant aux valeurs descriptives des instances. On obtient alors *K* sous-chaînes consécutives S_k de taille n_k (taille déterminée par la somme des

effectifs par valeurs descriptives constituant le groupe). On utilise à nouveau l'hypothèse d'indépendance entre les groupes. On obtient

$$\begin{aligned} p(S/M) &= p\left(S/K, \{k(i)\}, \{n_{kj}\}\right) \\ &= p\left(S_1, S_2, ..., S_K/K, \{k(i)\}, \{n_{kj}\}\right) \\ &= \prod_{k=1}^{K} p\left(S_k/K, \{k(i)\}, \{n_{kj}\}\right) \\ &= \prod_{k=1}^{K} p\left(S_k/\{n_{kj}\}_k\right) \\ &= \prod_{k=1}^{K} \frac{1}{\left(n_k!/n_{k,1}!n_{k,2}!...n_{k,J}!\right)}, \end{aligned}$$

car l'évaluation de la probabilité d'observer une sous-chaîne S_i avec un a priori uniforme correspond à la distribution multinomiale.

En prenant les opposés des logarithmes des probabilités, le problème de maximisation se transforme en problème de minimisation pour le critère du théorème

$$Value(M) = \log(I) + \log\left(B(I,K)\right) + \sum_{k=1}^{K} \log\left(C_{n_k+J-1}^{J-1}\right) + \sum_{k=1}^{K} \log\left(n_k!/n_{k,1}!n_{k,2}!...n_{k,J}!\right).$$

Théorème: Dans un modèle SGM optimal suivant un a priori à trois étages, deux valeurs descriptives mono-classe et de même classe sont nécessairement dans le même groupe.

Preuve:

Supposons au contraire qu'il y ait deux valeurs descriptives a et b mono-classe dans deux groupes différents.

Soient A1 et B1 les groupes contenant respectivement a et b. On suppose, sans perte de généralité, que la classe cible concernée est la classe d'index 1.

On peut construire deux groupes alternatifs A0 et B2 en déplaçant la valeur descriptive a de A1 vers B1.

Calculons la variation du coût de la solution en adoptant cette solution alternative.

$$\begin{aligned} \Delta Cost1 = \Delta PartitionCost \\ &+ \log\left((n_{A0}+J-1)!/n_{A0}!(J-1)!\right) + \log\left(n_{A0}!/n_{A0,1}!n_{A0,2}!...n_{A0,J}!\right) \\ &+ \log\left((n_{B2}+J-1)!/n_{B2}!(J-1)!\right) + \log\left(n_{B2}!/n_{B2,1}!n_{B2,2}!...n_{B2,J}!\right) \\ &- \log\left((n_{A1}+J-1)!/n_{A1}!(J-1)!\right) - \log\left(n_{A1}!/n_{A1,1}!n_{A1,2}!...n_{A1,J}!\right) \\ &- \log\left((n_{B1}+J-1)!/n_{B1}!(J-1)!\right) - \log\left(n_{B1}!/n_{B1,1}!n_{B1,2}!...n_{B1,J}!\right) \end{aligned}$$

Si A1 est réduit à a, ou B1 à b, il y a alors diminution du nombre de groupes, la variation de coût de partition est $\Delta PartitionCost = \log\left(B(I,K-1)\right) - \log\left(B(I,K)\right)$, donc négative. Sinon, le nombre de groupe étant constant, cette variation de coût est nulle.

Les effectifs par classe cible sont les mêmes, sauf pour la classe cible d'index 1.

$$\begin{aligned} \Delta Cost1 - \Delta PartitionCost &= \log\left((n_{A1}-n_a+J-1)!/(n_{A1}+J-1)!\right) + \log\left((n_{B1}+n_a+J-1)!/(n_{B1}+J-1)!\right) \\ &+ \log\left(n_{A1,1}!/(n_{A1,1}-n_{a,1})!\right) + \log\left(n_{B1,1}!/(n_{B1,1}+n_{a,1})!\right) \end{aligned}$$

$$\Delta Cost1 - \Delta PartitionCost = \log\left(\prod_{n=0}^{n_a-1}\left(n_{A1,1}-n\right)/\left(n_{A1}-n+J-1\right)\right) - \log\left(\prod_{n=1}^{n_a}\left(n_{B1,1}+n\right)/\left(n_{B1}+n+J-1\right)\right)$$

De même, on peut construire deux groupes alternatifs A2 et B0 en déplaçant la valeur descriptive b de B1 vers A1.

La variation du coût de la solution en adoptant cette seconde solution alternative est de façon similaire:

$$\Delta Cost2 - \Delta PartitionCost = \log\left(\prod_{n=0}^{n_b-1}\left(n_{B1,1}-n\right)/\left(n_{B1}-n+J-1\right)\right) - \log\left(\prod_{n=1}^{n_b}\left(n_{A1,1}+n\right)/\left(n_{A1}+n+J-1\right)\right)$$

Supposons que $n_{A1,1}/\left(n_{A1}+J-1\right) \le n_{B1,1}/\left(n_{B1}+J-1\right)$.

Pour $0 < x < y,\ 0 \le z \le x,\qquad (x-z)/(y-z) \le x/y \le (x+z)/(y+z)$

Cela entraîne:

$$\prod_{n=0}^{n_a-1}\left(n_{A1,1}-n\right)/\left(n_{A1}-n+J-1\right)\le\left(n_{A1,1}/n_{A1}+J-1\right)^{n_a}\le\left(n_{B1,1}/n_{B1}+J-1\right)^{n_a}\le\prod_{n=1}^{n_a}\left(n_{B1,1}+n\right)/\left(n_{B1}+n+J-1\right)$$

Alors $\Delta Cost1-\Delta PartitionCost\le 0$.

De même, si $n_{A1,1}/\left(n_{A1}+J-1\right)\ge n_{B1,1}/\left(n_{B1}+J-1\right)$, on obtient $\Delta Cost2-\Delta PartitionCost\le 0$.

On rappelle que dans tous les cas, $\Delta PartitionCost\le 0$. On a donc intérêt soit à faire passer la valeur descriptive a de A1 vers B1, soit à faire passer la valeur descriptive b de B1 vers A1. On a donc toujours intérêt à laisser les valeurs descriptives a et b dans le même groupe, dans la solution de coût minimal.

Remarque:

Si l'on évalue le coût des partitions par les nombres de Stirling de seconde espèce (a priori de modèle de groupage, avec partition des valeurs descriptives en *K* groupes non vides), ce théorème est invalide dans certains cas, par exemple lorsqu'un des groupes ne contient qu'une valeur descriptive et où le nombre de groupes *K* est proche du nombre *I* de modalités. En effet, dans ce type de cas, la diminution du nombre de groupe entraîne une augmentation du coût de partition (pour mémoire: $S(n,n-1)=n(n-1)/2$ et $S(n,n)=1$).

Théorème: Dans un modèle SGM suivant un a priori à trois étages dans le cas de deux classes cibles, s'il y a autant de valeurs descriptives que d'instances, alors le groupage optimal est réduit à un seul groupe contenant toutes les valeurs descriptives.

Preuve:

D'après le théorème précédent, toutes les valeurs descriptives constituent des singletons, donc sont nécessairement au moins groupées par classe. Il y a donc au plus *J* groupes, tous purs.

Soient deux groupes A et B, AuB le groupe résultant de la fusion de A et B.

Soient n_A, n_B et n_{AUB} les effectifs de ces groupes.

Soient $n_{A,j}$, $n_{B,j}$ et $n_{AUB,j}$ les effectifs des ces groupes par classe cible.

Calculons la variation du coût de groupage suite à la fusion de A et B, faisant passer le nombre de groupes de *K* à *K*-1.

$$\Delta Cost=\Delta PartitionCost+\left(\log\left(C_{n_{A\cup B}+J-1}^{J-1}\right)+\log\left(n_{A\cup B}!/n_{A\cup B,1}!n_{A\cup B,2}!...n_{A\cup B,J}!\right)\right)$$
$$-\left(\log\left(C_{n_A+J-1}^{J-1}\right)+\log\left(n_A!/n_{A,1}!n_{A,2}!...n_{A,J}!\right)\right)-\left(\log\left(C_{n_B+J-1}^{J-1}\right)+\log\left(n_B!/n_{B,1}!n_{B,2}!...n_{B,J}!\right)\right)$$

$$\Delta PartitionCost=\log\left(B\left(I,K-1\right)\right)-\log\left(B\left(I,K\right)\right)$$

$$\Delta Cost=\Delta PartitionCost+\log\left(\left(n_{A\cup B}+J-1\right)!\left(J-1\right)!/\left(n_A+J-1\right)!\left(n_B+J-1\right)!\right)-\sum_{j=1}^{J}\log\left(C_{n_{A\cup B,j}}^{n_{A,j}}\right)$$

On a ici la fusion de deux groupes contenant des classes complémentaires (une classe représentée dans un groupe est absente de l'autre groupe).

$$\Delta Cost=\Delta PartitionCost+\log\left(\left(n_{A\cup B}+J-1\right)!\left(J-1\right)!/\left(n_A+J-1\right)!\left(n_B+J-1\right)!\right)$$

On a ici $J=2$, $n=n_{AUB}$, $K=2$.

$$\Delta Cost=\log\left(B\left(n,1\right)\right)-\log\left(B\left(n,2\right)\right)+\log\left(\left(n+1\right)!/\left(n_A+1\right)!\left(n_B+1\right)!\right)$$

$$\Delta Cost=-\log\left(2^{n-1}\right)+\log\left(C_{n+2}^{n_A+1}\right)-\log(n+2)$$

Pour $n>1$ et $k>1$, on a $C_n^k=C_{n-1}^{k-1}+C_{n-1}^k\le 2^{n-1}$

En fait, l'inégalité est stricte dès que *n* est supérieur à 2.

$$\Delta Cost<-\log\left(2^{n-1}\right)+\log\left(2^{n+1}\right)-\log(n+2)$$

$$\Delta Cost<\log\left(4/\left(n+2\right)\right)$$

Donc, $\Delta Cost<0$.

Donc, le groupage optimal est réduit à un seul groupe contenant toutes les valeurs descriptives.

Remarque:

Si l'on évalue le coût des partitions par les nombres de Stirling de seconde espèce (a priori de modèle de groupage avec partition des valeurs descriptives en *K* groupes non vides), ce théorème est invalide dans le cas d'équidistribution des classes cibles. On vérifie en effet que dans ce cas, le coût de la partition en un seul groupe ($\log(n+1)+\log\left(C_n^{n/2}\right)$) est plus élevé que celui de la partition en *n* groupes ($n\log\left(2\right)$).

Conjecture: Dans un modèle SGM suivant un a priori à trois étages dans le cas de deux classes cibles, on peut ordonner les valeurs descriptives par proportion croissante de la première classe cible. Alors, toute valeur descriptive de proportion

comprise entre deux valeurs descriptives incluse dans un même groupe doit nécessairement figurer dans ce groupe dans le groupage optimal.

Quelques éléments:

Dans le cas de deux classes cibles, (Breiman 1984) a montré que la bi-partition optimale pouvait être trouvée en ordonnant préalablement les valeurs descriptives par proportion croissante de la première classe cible, puis en parcourant les valeurs descriptives dans cet ordre pour chercher la meilleure bipartition. Breiman a démontré cette propriété avec le critère de Gini utilisé dans l'algorithme CART. (Asseraf 1996) a montré la même propriété en utilisant cette fois la mesure de Kolmogorov-Smirnov. Dans le cas du critère d'évaluation des partitions MODL, cette propriété paraît également intuitive, mais semble difficile à démontrer, le critère étant basé sur des probabilités discrètes.

Afin d'évaluer la validité de la conjecture, des expérimentations numériques ont été menées de façon systématique sur tous les triplets de valeurs descriptives d'effectif inférieur à 100 et de distribution cible quelconque. Le groupage optimal a été calculé sur tous ces triplets en évaluant toutes les partitions possibles. Le nombre de cas évalués est de l'ordre de 10^{12}. Dans tous les cas, la partition optimale est compatible avec l'ordre des valeurs descriptives (induit par la proportion de la première classe cible). L'expérimentation a également été menée sur tous les quadruplets de valeurs descriptives d'effectif inférieur à 50 (nombre de cas évalués de l'ordre de 10^{12}) sans jamais invalider la conjecture. Le critère MODL étant additif sur l'ensemble des groupes constitués, la portée de cette validation numérique s'applique également aux sous-parties de partitions d'effectif plus important.

*4.3 Evaluation numérique du nombre de partitions B(*n,k*)*

Les bases de données réelles peuvent comporter des attributs symboliques ayant de très grands nombres de modalités: codes postaux, référence de produit, prénom… Il est alors nécessaire de pouvoir évaluer les nombres de partitions B(n,k) dans des domaines numériques inhabituels, nécessitant des méthodes numériques spécifiques.

4.3.1 Nombres de Stirling de deuxième espèce

$S(n,k)$ est le nombre de partition en k parties non vides d'un ensemble de n éléments. $S(n,k)$ est appelé le nombre de Stirling de deuxième espèce.

On a la relation de récurrence suivante: $S(n,k) = S(n-1,k-1) + kS(n-1,k)$ et la formule suivante:

$$S(n,k) = \frac{1}{k!}\sum_{i=0}^{k-1}(-1)^i C_k^i (k-i)^n .$$

$S(n,1) = S(n,n) = 1$.

$S(n,2) = 2^{n-1} - 1$.

$S(n,n-1) = n(n-1)/2$.

Pour k fixé, quand n tend vers l'infini, on a $S(n,k) \sim \frac{k^n}{k!}$.

L'évaluation numérique de $S(n,k)$ pose rapidement des problèmes. Si l'on utilise la formule basée sur une sommation de facteurs alternés, la précision numérique des ordinateurs (environ 15 digits) entraîne dès résultats erronés pour les valeurs k proches de n, dès que n dépasse 30. On doit effectivement additionner des facteurs alternés très grands, dont la somme globale peut être très petite ($S(n,n) = 1$). Si l'on utilise la relation de récurrence, on doit bufferiser les résultats pour avoir des temps de calcul raisonnables, ce qui limite le domaine de valeurs de n pour lequel on veut évaluer $S(n,k)$ (espace mémoire requis en n^2). De plus, on tombe rapidement sur une autre limite numérique qui est celle des exposants des nombres réels représentables sur ordinateur (environ 10^{300}). Dans ce cas, en utilisant la formule $S(n,k) \sim k^n/k!$, on peut vérifier que $S(250,40) > 10^{350}$, ce qui dépasse les limites informatiques. Il faut alors envisager d'évaluer le logarithme de $S(n,k)$, ce qui est de toute façon requis pour l'évaluation du critère MODL. Les formules additives permettant le calcul des $S(n,k)$ ne se prêtent pas à une évaluation logarithmique, ce qui invalide cette approche.

En conclusion, l'évaluation précise des $S(n,k)$ n'est pas envisageable numériquement au delà de n de l'ordre de 200, du moins avec les méthodes usuelles.

4.3.2 Evaluation numérique de B(n,k)

$B(n,k)$ est le nombre de partition en k parties (éventuellement vides) d'un ensemble de n éléments.

Pour k=n, ce nombre est égal au nombre total de partitions, ce qui correspond au nombre de Bell.

Par définition, on a $B(n,k) = \sum_{i=1}^{k} S(n,i)$.

$$B(n,k) = \sum_{i=1}^{k} \sum_{j=0}^{i-1} (-1)^j \frac{(i-j)^n}{j!(i-j)!}.$$

En regroupant les termes de type $(i-j)^n/(i-j)!$, on obtient:

$$B(n,k) = \sum_{i=1}^{k} \frac{i^n}{i!} \sum_{j=0}^{k-i} \frac{(-1)^j}{j!}.$$

Pour *k*=*n*, la formule ci-dessus se ramène à une variation de la formule de Dobinski pour les nombres de Bell.

On reconnaît le développement limité $e^{-1} = \sum_{j=0}^{\infty} \frac{(-1)^j}{j!}$, série convergeant très vite vers sa limite. $B(n,k)$ s'exprime alors sous la forme d'une somme de facteurs tous positifs, dont la plupart (sauf les derniers termes) peuvent se simplifier en $e^{-1} i^n/i!$.

Pour évaluer le logarithme de $B(n,k)$, on peut rechercher le terme dont la contribution est maximale. En utilisant la formule de Stirling pour approximer le terme factoriel, on obtient:

$i^n/i! \sim i^{n-i-1/2} e^i/\sqrt{2\pi}$.

La dérivée de cette fonction en *i* est $\left(i^{n-i-1/2} e^i/\sqrt{2\pi}\right)(n-1/2-i\ln(i)/i)$.

On a donc une fonction dont le maximum est atteint pour $i\ln(i) \sim n$, croissante avant ce maximum et décroissante après.

La fonction $r(x)$ définie par $r(x)e^{r(x)} = x$ a été étudiée dans la littérature (Canfield and Pomerance, 2002), et vérifie $\ln(x) - \ln(\ln(x)) \le r(x) \le \ln(x)$. Cela permet d'obtenir un intervalle de recherche restreint pour l'équation $i\ln(i) \sim n$, dont on pourrait trouver la solution entière i_{max} la plus proche par dichotomie. Pour l'évaluation de $B(n,k)$, l'indice i_0 correspondant à la valeur max de $i^n/i!$ vaut soit i_{max} si *k* est supérieur à i_{max}, soit *k* sinon (la fonction $i^n/i!$ est croissante avant i_{max}).

On peut alors réécrire $B(n,k)$ en:

$$B(n,k) = \frac{{i_0}^n}{i_0!} \sum_{i=1}^{k} \frac{i_0!}{{i_0}^n} \frac{i^n}{i!} e^{-1}(k-i) \text{ avec } e^{-1}(k-i) = \sum_{j=0}^{k-i} \frac{(-1)^j}{j!}.$$

Tous les facteurs de la somme sont compris entre 0 et 1 et peuvent être facilement évalué numériquement (en prenant l'exponentiel du logarithme des facteurs pour résoudre les problèmes d'estimation des puissances de *i* et des factorielles de *i*). Le total sous le signe somme est donc compris entre 1 et *k*, ce qui ne pose pas de problème numérique particulier.

Afin d'optimiser le temps de calcul, on peut commencer la sommation à partir de l'indice i_0 du facteur maximum, puis additionner les facteurs précédents et suivants (qui décroissent) tant que leur contribution est significative. Expérimentalement, on a vérifié que le nombre de termes significatifs à prendre en compte est de l'ordre de racine de *n*.

Lors d'expérimentations numériques, on a vérifié que pour *n*=200, la différence relative entre l'évaluation exacte par sommation des nombres de Stirling de seconde espèce (valide numériquement pour *n* <= 200) et l'évaluation par la méthode proposée ci-dessus est inférieure au seuil de précision informatique des réels (10^{-15}), et que cette méthode permet de calculer environ 100000 nombres de Bell généralisés par seconde sur un PC 1,7 Ghz.

En résumé, la méthode de calcul exposée ci-dessus permet une évaluation précise du logarithme de $B(n,k)$ dans des domaines numériques très au delà de $n=10^9$, ce qui permet une évaluation précise du critère MODL sans être limité par la taille des bases de données.

4.3.3 Tables numériques

Tableau 3 : Nombres de Stirling de seconde espèce $S(n,k)$

n \ k	1	2	3	4	5	6	7	8	9	10
1	1									
2	1	1								
3	1	3	1							
4	1	7	6	1						
5	1	15	25	10	1					
6	1	31	90	65	15	1				
7	1	63	301	350	140	21	1			
8	1	127	966	1701	1050	266	28	1		
9	1	255	3025	7770	6951	2646	462	36	1	
10	1	511	9330	34105	42525	22827	5880	750	45	1
11	1	1023	28501	145750	246730	179487	63987	11880	1155	55
12	1	2047	86526	611501	1379400	1323652	627396	159027	22275	1705

Tableau 4 : Nombres de Bell généralisés $B(n,k)$

n \ k	1	2	3	4	5	6	7	8	9	10
1	1									
2	1	2								
3	1	4	5							
4	1	8	14	15						
5	1	16	41	51	52					
6	1	32	122	187	202	203				
7	1	64	365	715	855	876	877			
8	1	128	1094	2795	3845	4111	4139	4140		
9	1	256	3281	11051	18002	20648	21110	21146	21147	
10	1	512	9842	43947	86472	109299	115179	115929	115974	115975
11	1	1024	29525	175275	422005	601492	665479	677359	678514	678569
12	1	2048	88574	700075	2079475	3403127	4030523	4189550	4211825	4213530