NT/FTR&D/8028

10 Mars 2003

DTL
Direction des
Techniques
Logicielles

Groupage robuste des valeurs d'un attribut symbolique par la méthode Khiops

Marc Boullé (DTL/TIC)

Note Technique

france telecom
R&D

Note Technique
NT/FTR&D/8028

10 mars 2003

Groupage robuste des valeurs d'un attribut symbolique par la méthode Khiops

Marc Boullé (DTL/TIC)

Vu, pour accord le
directeur de DTL

J.M. Pitié

Vu, le responsable du
laboratoire TIC

O. Perrault

Date : 10 mars 2003

Résumé : Dans le domaine de l'apprentissage supervisé, les méthodes de groupage des modalités d'un attribut symbolique permettent de construire un nouvel attribut synthétique conservant au maximum la valeur informationnelle de l'attribut initial et diminuant le nombre de modalités. Nous proposons ici une généralisation de l'algorithme de discrétisation Khiops pour le problème du groupage des modalités. L'algorithme proposé permet de contrôler a priori le risque de sur-apprentissage et d'améliorer significativement la robustesse des groupages produits. Cette caractéristique de robustesse a été obtenue en étudiant la statistique des variations du critère du Khi2 lors de regroupements de lignes d'un tableau de contingence et en modélisant le comportement statistique de l'algorithme Khiops. Cette modélisation, vérifiée expérimentalement, permet d'apporter des garanties concernant le rique de sur-apprentissage. Des expérimentations intensives ont été menées sur de nombre jeux d'essai synthétiques et réels afin de comparer la méthode Khiops avec d'autres méthodes de groupage.

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

France Télécom R&D
Branche 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 4 760 634 896 € - 380 129 866 RCS Paris

Groupage robuste des valeurs d'un attribut symbolique par l'algorithme Khiops

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 modalités d'un attribut symbolique permettent de construire un nouvel attribut synthétique conservant au maximum la valeur informationnelle de l'attribut initial et diminuant le nombre de modalités. Nous proposons ici une généralisation de l'algorithme de discrétisation Khiops[*] pour le problème du groupage des modalités. L'algorithme proposé permet de contrôler a priori le risque de sur-apprentissage et d'améliorer significativement la robustesse des groupages produits. Cette caractéristique de robustesse a été obtenue en étudiant la statistique des variations du critère du Khi2 lors de regroupements de lignes d'un tableau de contingence et en modélisant le comportement statistique de l'algorithme Khiops. Cette modélisation, vérifiée expérimentalement, permet d'apporter des garanties concernant le risque sur-apprentissage. Des expérimentations intensives ont été menées sur de nombre jeux d'essai synthétiques et réels afin de comparer la méthode Khiops avec d'autres méthodes de groupage.

Mots clés : analyse intelligente donnée ; apprentissage automatique ; groupage

[*] Dépôts de brevet N° 01 07006 et N° 02 16733

TABLE DES MATIERES

1 Introduction 3
2 Méthode de groupage Khiops initial 4
2.1 Le test du Khi2 : principes et notations 4
2.2 Méthode de discrétisation Khiops initiale 5
2.3 Méthode de groupage Khiops initiale 5
3 Méthode de groupage Khiops robuste 6
3.1 Principe de l'amélioration de la robustesse de l'algorithme 6
3.2 Variation du Khi2 suite à la fusion de deux lignes du tableau de contingence 6
3.3 Statistique du MaxDeltaKhi2 de groupage 7
3.4 Amélioration de la gestion de la contrainte d'effectifs minimum 8
3.5 Amélioration de la complexité de l'algorithme 8
4 Evaluation 8
4.1 Méthode d'évaluation du groupage 8
4.2 Méthodes évaluées 9
4.2.1 Méthode Khiops initiale 9
4.2.2 Méthode Khiops 9
4.2.3 Méthode Tschuprow 9
4.2.4 Méthode CHAID 10
4.2.5 Méthode Gain Ratio 10
4.3 Groupage d'un attribut descriptif indépendant de l'attribut à prédire 11
4.4 Jeux d'essai théoriques 12
4.5 Benchmarks UCI 13
4.5.1 Qualité des groupages 14
4.5.2 Taille des groupages 15
4.5.3 Synthèse des résultats 16
Conclusion 17
Références 17
5 Annexe : étude de la statistique du MaxDeltaKhi2 de groupage 18
5.1 Présentation 18
5.2 Insensibilité à la taille de l'échantillon 18
5.3 Insensibilité à la répartition des modalités cibles 18
5.4 Insensibilité à la répartition des modalités descriptives 19
5.5 Cas avec deux modalités descriptives 20
5.6 Cas avec deux modalités cibles 20
5.7 Cas général 21
5.8 Approximation de la loi de MaxDeltaKhi2 par la loi normale 22
5.9 Remarque 23

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 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'arbre 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 variables 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, de la régression linéaire 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 du groupage à optimiser. Plusieurs stratégies de groupage ont été explorées dans la bibliographie. L'algorithme le plus simple est la binarisation en choisissant d'isoler une modalité contre toutes les autres. Une stratégie plus élaborée consiste à rechercher le regroupement optimal des modalités en deux groupes (optimal au sens du critère mathématique 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 modalités 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 d'individus associés à la première classe, 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 modalités descriptives originelles. En se basant sur les idées de (Lechevallier 1990; Fulton, Kasif & Salzberg 1995), il paraît 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 modalités descriptives originelles. 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 modalités 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 proche 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 variable 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 modalités 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 modalités 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 modalités 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. Pour un overview complet sur les méthodes à base d'arbre ou de graphe d'induction et la façon dont elles traitent les variables 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'individus 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, et un algorithme conduisant à la meilleure partition.

La méthode de groupage Khiops initiale est une généralisation directe de la méthode de discrétisation Khiops (Boullé 2002). Cette méthode utilise la valeur globale du Khi2 du tableau de contingence entre attribut discrétisé et attribut à prédire, et cherche à minimiser la probabilité d'indépendance correspondante. La méthode de groupage initialise le partitionnement à partir des modalités descriptives originelles, évalue toutes les fusions possibles et choisit celle qui maximise le critère du Khi2 appliqué à la nouvelle partition formée. La méthode s'arrête automatiquement dès que la probabilité d'indépendance entre le groupage et les classes cibles ne décroît plus. Cet algorithme de groupage est amélioré en se basant sur les idées mises en œuvre pour la version robuste de l'algorithme de discrétisation Khiops (Boullé 2002). Cette version robuste permet un contrôle réel de la qualité prédictive d'un groupage de modalités. Le principe utilisé repose sur l'étude du comportement statistique de l'algorithme en présence d'un attribut symbolique indépendant de l'attribut à prédire. Nous avons étudié la statistique de la variation maximale du critère du Khi2 lors du déroulement complet de l'algorithme de groupage. Cette étude a montré que ce critère MaxDeltaKhi2 est insensible à la répartition des modalités et à la taille de l'échantillon d'apprentissage, et ne dépend que du nombre de modalités des attributs descriptifs et cibles. Suite à cette modélisation de la statistique du MaxDeltaKhi2, l'algorithme de groupage initial a été modifié en le contraignant à accepter toute fusion de modalités entraînant une variation du Khi2 inférieure à la variation théorique maximale calculée. Cette amélioration permet de garantir d'une part que les groupages de modalités d'un attribut indépendant de l'attribut à prédire aboutissent à un seul groupe terminal, d'autre part que les groupages aboutissant à plusieurs groupes correspondent à des attributs ayant un intérêt prédictif réel. Des expérimentations confirment l'intérêt de cette version robuste de l'algorithme Khiops et montrent de bonnes performances prédictives pour les groupages obtenus.

Le document est organisé de la façon suivante.

La partie 2 rappelle l'algorithme de discrétisation Khiops initial et en présente la généralisation au groupage. La partie 3 présente l'amélioration de la robustesse de la méthode de groupage Khiops, basée sur la connaissance de la statistique de l'algorithme. La partie 4 procède à des expérimentations intensives sur des bases synthétiques et réelles. L'annexe détaille l'étude de la statistique du MaxDeltaKhi2 de l'algorithme de groupage.

2 Méthode de groupage Khiops initial

2.1 Le test du Khi2 : principes et notations

Il s'agit de tester l'hypothèse d'indépendance entre un attribut descriptif et un attribut cible. Dans un premier temps, toutes les instances du jeu de données sont résumées dans un tableau de contingence, qui contient pour chaque paire de valeurs descriptive et cible la fréquence (nombre d'instances) correspondante. La valeur du Khi2 est calculée à partir du tableau de contingence, en se basant sur les notations présentées dans le tableau 1.

Tableau 1 : Tableau de contingence utilisée pour le calcul de la valeur du Khi2

n_{ij}: Fréquence observée pour la $i^{ème}$ valeur descriptive
Et la $j^{ème}$ valeur cible
$n_{i.}$: Fréquence totale observée pour la $i^{ème}$ valeur descriptive
$n_{.j}$: Fréquence totale observée pour la $j^{ème}$ valeur cible
N: Fréquence totale observée
I: Nombre de valeurs descriptives
J: Nombre de valeurs cibles

	A	B	C	Total
a	n_{11}	n_{12}	n_{13}	$n_{1.}$
b	n_{21}	n_{22}	n_{23}	$n_{2.}$
c	n_{31}	n_{32}	n_{33}	$n_{3.}$
d	n_{41}	n_{42}	n_{43}	$n_{4.}$
e	n_{51}	n_{52}	n_{53}	$n_{5.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	N

Soit $e_{ij} = n_{i.} * n_{.j} / N$, la fréquence attendue pour la cellule (i, j) du tableau de contingence dans le cas où les attributs descriptif et cible sont indépendants. La valeur du Khi2 est une mesure sur l'ensemble du tableau de contingence de la différence entre les fréquences observées et les fréquences attendues. Elle peut être interprétée comme une distance à l'hypothèse d'indépendance entre les attributs.

$$Khi2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

Sous l'hypothèse nulle d'indépendance, la valeur du Khi2 suit une loi du Khi2 à (I-1)*(J-1) degrés de liberté. Ceci constitue le fondement d'un test statistique permettant de rejeter l'hypothèse d'indépendance. Plus la valeur du Khi2 est importante, moins l'hypothèse d'indépendance est probable.

2.2 Méthode de discrétisation Khiops initiale

Le test du Khi2 est à la fois sensible aux effectifs et aux proportions des classes cibles, à la fois localement pour chaque ligne de la table de contingence et globalement pour l'ensemble de la table. Il s'agit donc d'un critère intéressant a priori pour les méthodes de discrétisation ou de groupage. La loi du Khi2 dépend du nombre de modalités (par le paramétrage du nombre de degrés de liberté). Cependant, en passant de la valeur du Khi2 à la valeur de la probabilité d'indépendance associée, on peut comparer deux discrétisations basées sur des nombres d'intervalles différents.

On cherche à minimiser la probabilité d'indépendance entre la loi discrétisée et la loi cible en passant par la loi du Khi2. Les conditions d'application du test du Khi2 imposent que l'on ait un effectif théorique minimum de 5 dans chaque cellule du tableau de Khi2. Afin de se prémunir (de façon heuristique) contre le sur-apprentissage, cet effectif minimum est renforcé, en imposant que chaque intervalle ait un effectif au moins égal à la racine carrée de l'échantillon. Ces contraintes sont prises en compte dans l'algorithme d'optimisation de la discrétisation.

La méthode d'optimisation utilisée est une méthode gloutonne de type ascendante. On part des intervalles élémentaires, et on recherche la meilleure fusion possible, c'est à dire celle qui entraîne en priorité un meilleur respect des contraintes d'effectif minimum, et à respect de contrainte égal, celle qui minimise la probabilité d'indépendance entre attribut discrétisé et attribut cible. On s'arrête quand toutes les contraintes sont respectées et qu'aucune fusion supplémentaire ne diminue la probabilité d'indépendance entre attribut discrétisé et attribut cible.

Algorithme Khiops initial

- Initialisation
  - Tri des valeurs de l'attribut descriptif
  - Création d'un intervalle élémentaire par valeur de l'attribut descriptif
  - Calcul de la probabilité d'indépendance entre l'attribut discrétisé et l'attribut cible
- Optimisation de la discrétisation
  - Répéter
    - Evaluer toutes les fusions possibles d'intervalles adjacents
      - ✓ Calcul du Khi2 associé à la nouvelle discrétisation résultant de la fusion
    - Chercher la meilleure fusion
      - ✓ Fusions améliorant le respect des contraintes en priorité
      - ✓ Maximum du Khi2
    - Evaluer la condition d'arrêt
      - ✓ Arrêter si toutes les contraintes sont respectées ou si la probabilité d'indépendance augmente suite à la fusion
      - ✓ Continuer sinon (et effectuer la meilleure fusion)

A la fin de l'algorithme, on définit un indicateur de qualité de la discrétisation *ProbLevel* en se basant sur la probabilité d'indépendance entre l'attribut discrétisé et l'attribut cible.

$ProbLevel = -\log_{10}\left(P\left(Khi2_{final}\right)\right)$ (si discrétisation à plusieurs intervalles, 0 sinon).

L'évaluation de la probabilité d'indépendance pose des problèmes numériques quand la taille du tableau de contingence est importante où quand l'attribut cible et l'attribut à prédire sont très corrélés sur un échantillon de taille importante. Ces problèmes ont été étudiés et résolus dans (Boullé 2002), en se basant sur une bonne approximation du logarithme de la probabilité d'indépendance, et pour une meilleure précision sur une évaluation précise du seuil de variation de la valeur du Khi2 contrôlant le critère d'arrêt de l'algorithme.

La complexité de l'algorithme est en N^3 avec une implémentation naïve. En se basant sur la bufferisation des calculs du Khi2 et de ses variations, et sur l'utilisation d'une liste triée des fusions d'intervalles, il est possible de ramener la complexité de l'algorithme à N.log(N). Cette version optimisée de l'algorithme est décrite dans (Boullé 2002).

2.3 Méthode de groupage Khiops initiale

L'algorithme de discrétisation se généralise au groupage en remplaçant les intervalles par des groupes de modalités, et en remplaçant la recherche de la meilleure fusion d'intervalles adjacents par la recherche de la meilleure fusion de groupes quelconques. La contrainte d'effectif minimum par intervalle se traduit ici par un effectif minimum par modalité. Lors d'un prétraitement, toute modalité descriptive originelle n'atteignant pas cet effectif minimum sera groupée inconditionnellement vers une modalité spéciale AUTRE. Il ne reste alors que des modalités satisfaisant la contrainte d'effectif minimum en entrée de l'algorithme de groupage.

Algorithme Khiops initial de groupage:

- Initialisation
  - Tri des valeurs de l'attribut descriptif
  - Initialisation des modalités descriptives originelles
  - Création de la modalité spéciale AUTRE, fusion de toutes les modalités originelle n'atteignant pas l'effectif

minimum
- Calcul de la probabilité d'indépendance entre l'attribut groupé et l'attribut cible

- Optimisation du groupage : répéter
  - Evaluer toutes les fusions inter groupes
    - ✓ Calcul du Khi2 associé au nouveau groupage résultant de la fusion
  - Chercher la meilleure fusion
    - ✓ Fusion maximisant la valeur du Khi2
  - Evaluer la condition d'arrêt
    - ✓ Arrêter si la probabilité d'indépendance diminue suite à la fusion
    - ✓ Continuer sinon (et effectuer la meilleure fusion)

De façon analogue à l'algorithme de discrétisation, il est possible de ramener l'algorithme de groupage à une complexité algorithmique de $N\log(N)+I^2\log(I)$ où N est le nombre d'individus de l'échantillon et I est le nombre de modalités de l'attribut descriptif (une fois la modalité spéciale AUTRE traitée).

3 Méthode de groupage Khiops robuste

3.1 Principe de l'amélioration de la robustesse de l'algorithme

Lors du déroulement de l'algorithme de groupage, toutes les fusions de lignes du tableau de contingence sont envisagées, et on choisit celle qui maximise la valeur du Khi2 du tableau de contingence après fusion des lignes, c'est à dire celle qui minimise la variation DeltaKhi2 du Khi2 au cours de la fusion. Le MaxDeltaKhi2 est la valeur maximale du DeltaKhi2 atteinte lors du déroulement complet de l'algorithme, c'est à dire jusqu'à l'obtention d'un unique groupe de modalités terminal. Le principe de la version robuste de l'algorithme est de constater que pour un attribut descriptif indépendant de l'attribut à prédire, on observe des variations stochastiques du DeltaKhi2, bornées par une valeur MaxDeltaKhi2 sur l'ensemble des étapes de l'algorithme. Un objectif naturel d'une version robuste de l'algorithme est d'assurer que le groupage des modalités d'un attribut indépendant de l'attribut à prédire aboutit à un seul groupe terminal. Pour cela, il suffit d'imposer que toute fusion de groupe entraînant une variation du Khi2 inférieure aux variations pouvant être dues au hasard (c'est à dire inférieure au MaxDeltaKhi2) soit automatiquement acceptée. On assure ainsi également que tout groupage aboutissant à plusieurs groupes terminaux correspond à un attribut non indépendant de l'attribut à prédire. De cette façon, on améliore significativement la robustesse de l'algorithme en apportant certaines garanties de résultat.

Algorithme Khiops robuste de groupage:
- Initialisation
  - Tri des valeurs de l'attribut descriptif
  - Initialisation des modalités descriptives originelles
  - Création de la modalité spéciale AUTRE, fusion de toutes les modalités originelle ne respectant pas la contrainte d'effectif minimum
  - Calcul de la probabilité d'indépendance entre attribut groupé et attribut cible
  - *Calcul de la valeur du MaxDeltaKhi2*
- Optimisation du groupage : répéter
  - Evaluer toutes les fusions inter groupes
    - ✓ Calcul du Khi2 associé à la nouvelle loi groupée résultant de la fusion
  - Chercher la meilleure fusion
    - ✓ Fusion maximisant la valeur du Khi2
  - Evaluer la condition d'arrêt
    - ✓ *Si le meilleur DeltaKhi2 est inférieur à MaxDeltaKhi2, continuer*
    - ✓ Sinon, si la probabilité d'indépendance diminue suite à la fusion, continuer
    - ✓ Sinon, arrêter
  - Si continuer, effectuer la meilleure fusion

Il reste maintenant à calculer la valeur du MaxDeltaKhi2 pour le problème du groupage.

3.2 Variation du Khi2 suite à la fusion de deux lignes du tableau de contingence

Evaluons la variation de la valeur du Khi2 suite au regroupement de deux lignes du tableau du Khi2, d'effectifs n et n', pour des proportions de modalités cibles locale p_j et p'_j, ces proportions étant P_j sur l'ensemble du tableau de contingence.

					Total
	…	…	…	…	…
	…	…	…	…	
	$p_1 n$	$p_2 n$	…	$p_1 n$	n
	$p'_1 n$	$p'_2 n$	…	$p'_1 n'$	n'
	…	…	…	…	
	…	…	…	…	
Total	$P_1 N$	$P_2 N$	…	$P_J N$	N

$$Khi2_{apèsfusion} - Khi2_{avantfusion} = -\frac{nn'}{n+n'}\sum_{j=1}^{J}\frac{(p_j - p'_j)^2}{P_j}$$

Cette variation est toujours négative, et n'est nulle que si les lignes ont exactement les mêmes proportions de modalités cibles. Le Khi2 d'un tableau de contingence ne peut que décroître suite à la fusion de deux lignes du tableau. Par la suite, on redéfinit le DeltaKhi2 en valeur absolue pour ne manipuler que des grandeurs positives.

$$DeltaKhi2 = \frac{nn'}{n+n'}\sum_{j=1}^{J}\frac{(p_j - p'_j)^2}{P_j}$$

On a montré dans (Boullé 2002) que dans le cas d'un attribut descriptif indépendant d'un attribut cible à J degrés de libertés, le DeltaKhi2 de la fusion de deux lignes de même effectif suit une loi du Khi2 à J-1 degrés de libertés.

3.3 Statistique du MaxDeltaKhi2 de groupage

Soit N la taille de l'échantillon, I le nombre de modalités descriptives et J le nombre de modalités cibles. On se place dans le cas où la contrainte d'effectif minimum de 5 par cellule du tableau de contingence est respectée, de façon à pouvoir utiliser la statistique du Khi2. Le MaxDeltaKhi2 est défini comme étant la variation maximale du Khi2 lors du déroulement complet de l'algorithme de regroupement jusqu'à un seul groupe terminal, en présence d'un attribut descriptif indépendant de l'attribut cible. A priori, la statistique du MaxDeltaKhi2 dépend de la taille de l'échantillon N, du nombre de modalités descriptives I, du nombre de modalités cibles J, de la répartition des fréquences des modalités descriptives et de la répartition des fréquences des modalités cibles. On montre en annexe que la loi du MaxDeltaKhi2 ne dépend en fait que de I et J, c'est à dire des nombres de modalités descriptives et cibles. De façon plus précise, on a obtenu les résultats suivants, la plupart de façon expérimentale.

Proposition 1: Pour 2 modalités descriptives et J modalités cibles, la loi du MaxDeltaKhi2 est la loi du Khi2 à J-1 degrés de libertés. Sa moyenne est donc J-1.

Dans le cas limite à 2 modalités descriptives, le MaxDeltaKhi2 se ramène en effet à la loi du Khi2, car l'algorithme de groupage se réduit à une unique fusion.

Proposition 2: Pour I modalités descriptives équidistribuées et 2 modalités cibles équidistribuées, la valeur moyenne du MaxDeltaKhi2 est asymptotiquement proportionnelle à $2I/\pi$.

Le principe de la preuve présentée en annexe est que l'on obtient la valeur maximum du DeltaKhi2 en regroupant dans un premier groupe toutes les modalités descriptives pour lesquelles la première modalité cible est majoritaire, et dans un second groupes les autres modalités descriptives. Il s'agit alors d'évaluer les proportions de modalités cible dans chacun de ces deux groupes, afin d'évaluer le MaxDeltaKhi2 correspondant.

Les propositions suivantes sont des conjonctures, et n'ont été évaluées qu'expérimentalement. Elles sont néanmoins intuitivement fondées sur la formule de calcul du Khi2, qui normalise les différences de fréquences par les effectifs attendus, ce qui rend la valeur moyenne du Khi2 et du MaxDeltaKhi2 indépendantes des répartitions des fréquences des modalités.

- La loi du MaxDeltaKhi2 ne dépend pas de l'effectif de l'échantillon.
- La loi du MaxDeltaKhi2 ne dépend pas de la répartition des fréquences des modalités descriptives.
- La loi du MaxDeltaKhi2 ne dépend pas de la répartition des fréquences des modalités cibles.

Dans le cas général, ne sachant pas décrire analytiquement la loi de MaxDeltaKhi2, on en a calculé la moyenne et l'écart type empiriquement en se basant sur 1000 expérimentations de groupage pour un grand nombre de couples de paramètres (I,J). Le tracé des courbes des moyennes et écart type fait apparaître un comportement quasiment linéaire vis à vis des paramètres I et J dès que ceux-ci dépassent quelques unités. Cette constatation justifie l'utilisation d'une table de valeurs, puis d'interpolations linéaires pour approximer les valeurs du MaxDeltaKhi2. En ce qui concerne la fonction de répartition du MaxDeltaKhi2, on a fait l'hypothèse confirmée par des expérimentations qu'elle se comportait de façon analogue à une loi normale de même moyenne et écart type.

En définitive, le MaxDeltaKhi2 utilisé pour l'algorithme de groupage est déterminé en recherchant par interpolation linéaire dans une table de valeurs préalablement calculée la valeur de la moyenne et de l'écart type du MaxDeltaKhi2 correspondant au nombre de modalités descriptives et au nombre de modalités cibles, puis en utilisant la loi normale inverse pour estimer la valeur du MaxDeltaKhi2 qui ne sera pas dépassée dans 95% des cas.

La méthodologie utilisée pour rendre robuste l'algorithme de groupage Khiops est généralisable à d'autres méthodes de

groupage, d'autant plus qu'elle repose pour l'essentiel sur une étude expérimentale de la statistique de l'algorithme. Cependant, les "bonnes" propriétés du critère du Khi2 permettent une réduction drastique des paramètres de cette statistique à seulement deux paramètres, à savoir les nombres de modalités descriptives et cibles. L'application de cette méthodologie ne paraît alors pas réaliste pour des critères de groupage ne permettant pas une telle simplification des paramètres de la statistique de l'algorithme.

3.4 Amélioration de la gestion de la contrainte d'effectifs minimum

Quand les modalités cibles ont des fréquences très déséquilibrées (par exemple 1% et 99%), la contrainte d'effectif minimum par ligne devient très importante (dans l'exemple, il faut au moins un effectif de 500 par modalité descriptive originelle). Si cela se comprend pour des modalités descriptives proches de la répartition globale (0-500, 5-500, 10-500 par exemple), cette contrainte paraît trop forte dans le cas d'un déséquilibre inverse très important (100-0 par exemple). Pour ne pas perdre inutilement des modalités potentiellement informatives en les « noyant » dans la modalité AUTRE, il est nécessaire de relâcher cette contrainte en trouvant un compromis entre la fiabilité du test du Khi2, qui est fondamentale car à la base de l'algorithme Khiops, et la finesse potentielle de groupage. Selon certains auteurs, l'effectif minimum théorique est de 5 par cellule du tableau de contingence, mais peut être ramené à 3, voire à 1 pour une seule classe en queue de distribution. On va alors relâcher la contrainte en imposant un effectif minimum par groupe de 6 (ce qui correspond à deux modalités cibles équidistribuées et un effectif minimum théorique de 3). Cet effectif minimum est très faible et donc risqué pour une évaluation fiable de la statistique du Khi2. On va renforcer cet effectif en se basant sur le calcul de la probabilité d'apparition d'une modalité descriptive pure vis à vis des classes cibles. Dans le cas de I modalités descriptives et deux modalités cibles équidistribuées, la probabilité d'observer au moins une modalité descriptive pure d'effectif k est $p_{k,I} \sim 1-(1-(1/2)^k)^I \sim I(1/2)^k$. Pour une probabilité p donnée de ne pas grouper un attribut indépendant de l'attribut cible, il est donc légitime que l'effectif minimum par groupe N_{min} soit au moins de la longueur k correspondante à $p=1-p_{k,M}$. Donc $N_{min} >= \log(1-p^{1/I})/\log(1/2)$. On peut vérifier que cette inégalité peut être approximée par $N_{min} >= \log_2(I)-\log_2(1-p)$, et donc que l'effectif minimum par groupe correspondant croît de façon logarithmique avec le nombre de modalités descriptives et avec le niveau de confiance demandé.

En résumé, on choisit un effectif minimum dépendant du nombre de modalités descriptives et du paramètre p de l'algorithme Khiops en imposant les contraintes suivantes :

- Effectif minimum de 3 par case du tableau de contingence correspondant à deux modalités cibles équidistribuées
  - $N_{min} >= 6$
- Pour une probabilité p donnée, taille minimum d'une modalité descriptive pure pour un attribut cible équidistribué
  - $N_{min} >= \log(1-p^{1/I})/\log(1/2)$

De cette façon, les effectifs minimums pour la statistique du Khi2 sont réduits au minimum pour permettre de détecter des petits groupes pertinents, et la contrainte exploite le paramètre p et la taille de l'ensemble d'apprentissage pour renforcer la fiabilité de l'évaluation du Khi2 dès que possible.

Le compromis entre fiabilité du critère du Khi2 et finesse de groupage est une limite de l'algorithme. La solution optimale à ce problème serait d'utiliser non pas la statistique du Khi2 dont la validité est asymptotique, mais la statistique exacte de Fischer. Le calcul de cette statistique discrète est combinatoire, ce qui rend cette solution optimale inutilisable en pratique.

3.5 Amélioration de la complexité de l'algorithme

Quand il y a de très nombreuses modalités originelles, la complexité algorithmique est de l'ordre de N^2, ce qui est trop grand. Dans le cas particulier où il n'y a que deux modalités cibles, on peut s'inspirer de l'algorithme de bipartition optimal de (Breiman, Friedman, Olshen et Stone 1984) en classant les modalités descriptives par proportion croissante de la première modalité cible et en n'envisageant que les regroupements de modalités adjacentes selon cet ordonnancement. On se ramène alors à une complexité équivalente au cas de la discrétisation, c'est à dire à Nlog(N). Dans le cas général où il y a au moins trois modalités cibles, on peut contrôler la complexité algorithmique et la limiter à une complexité de l'ordre de Nlog(N) en continuant dans la phase de prétraitement la fusion inconditionnelle des modalités descriptives vers la modalité AUTRE jusqu'à obtenir $\sqrt{N}$ modalités descriptives, et donc une complexité de l'algorithme en Nlog(N). Ces améliorations de la complexité de l'algorithme n'ont pas été implémentées pour les évaluations.

4 Evaluation

4.1 Méthode d'évaluation du groupage

L'enjeu 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 originelles. Les méthodes de prétraitement des attributs, discrétisation et groupage, sont généralement évaluées au moyen d'un algorithme inductif, en général un arbre de décision ou un prédicteur bayésien 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 masqué 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 de prédire chaque classe cible pour 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 modalités descriptives originelles 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 modalité 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, et d'autre part le prédicteur majoritaire, qui prédit systématiquement la classe majoritaire de l'échantillon en groupant toutes les modalités 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 inutilisable 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 modalités 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 distance 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 modalités descriptives en test.

Pour une modalité descriptive donnée, soit p_j la probabilité de la $j^{ème}$ modalité cible, estimée sur l'ensemble de test en se basant sur toutes les instances associées à la modalité descriptive, et soit q_j la probabilité de la $j^{ème}$ modalité cible, estimée sur l'ensemble d'apprentissage en se basant sur toutes les instances associées au groupe contenant la modalité descriptive. Alors la distance de Kullback-Leibler entre la distribution observée et la distribution estimée est définie par

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

Les probabilités p_j et q_j sont estimées par l'estimateur de Laplace pour des raisons de lissage et de gestion des probabilités nulles. La distance globale est calculée en prenant la moyenne des distances sur l'ensemble des instances de l'échantillon de test.

4.2 Méthodes évaluées

Nous allons évaluer la méthode Khiops en la comparant avec d'autres méthodes basées également 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 le même algorithme glouton de fusions itératives des modalités originelles, et ne seront donc différenciées que sur le critère utilisé.

4.2.1 Méthode Khiops initiale

La méthode Khiops initiale minimise la probabilité d'indépendance entre l'attribut groupé et l'attribut à prédire. La méthode ne nécessite aucun paramétrage, mais ne pourra produire que des partitions ayant au moins deux groupes.

En prétraitement, toutes les valeurs descriptives n'atteignant pas un effectif minimum sont regroupées dans une modalité spéciale. Cet effectif minimum garantit la fiabilité du test du Khi2 en imposant un effectif d'au moins 5 par cellule du tableau de contingence, et essaye de se prémunir du sur-apprentissage en imposant un effectif minimum par valeur descriptive au moins égal à la racine carrée de la taille de l'échantillon.

4.2.2 Méthode Khiops

La méthode Khiops 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. Cette effectif minimum est calculé au plus juste pour permettre un compromis entre fiabilité du test du Khi2 et finesse de la partition produite.

4.2.3 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. Avec les notations du tableau 1, 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 en fait 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 modalités 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.

4.2.4 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 modalités descriptives originelles, la méthode CHAID est identique à la méthode Khiops (exceptée la gestion de l'effectif minimum par groupe dans la phase de prétraitement).

4.2.5 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 modalités descriptives mono-classes 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 originelle. 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 originelles 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 modalités descriptives originelles 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 les méthodes Khiops initiale et 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 ne peut 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 $I^2\log(I)$ où I est le nombre de valeurs descriptives originelles. Dans le cas de la méthode Gain Ratio, cette

complexité est au moins en I^3.

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

Dans cette expérimentation, on groupe un attribut symbolique indépendant de l'attribut à prédire, pour diverses tailles d'échantillon, de nombres de modalités descriptives, de nombres de modalités cibles. On a également fait varier la répartition des modalités descriptives ou cibles en contrôlant un taux de déséquilibre des fréquences. Ce taux est égal au rapport entre la fréquence la plus importante et la fréquence la plus faible, l'ensemble des fréquences suivant une loi géométrique.

L'expérimentation est menée 1000 fois sur des échantillons générés aléatoirement pour chaque type de jeux d'essai. La taille des groupages est mesurée simplement en comptant le nombre de groupes produits. La qualité des partitions est évaluée en utilisant la distance 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 distance 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.

Afin d'améliorer la lisibilité, la valeur de la distance de Kullback-Leibler a été normalisée par la valeur correspondante obtenue dans le cas sans groupage (pour lequel la distribution prédite est obtenue par observation des fréquences cible sur la valeur descriptive de l'ensemble d'apprentissage). Le tableau 2 présente l'ensemble des résultats des expérimentations obtenues avec les méthodes testées. Les caractéristiques des benchmarks sont décrites dans les colonnes "Nb" (nombre de modalités) et "Rep." (répartition: taux de déséquilibre des fréquences des modalités) pour l'attribut descriptif et l'attribut cible. Les résultats sont résumés de façon synthétique sur la dernière ligne du tableau, en prenant leur moyenne (moyenne géométrique pour la distance de Kullback-Leibler, qui est sujette à des variations d'échelle importantes).

Tableau 2 : Nombre de groupes et distance de Kullback-Leibler des méthodes de groupage testées sur vingt familles de jeux d'essai dans le cas d'indépendance entre attribut descriptif et attribut cible

Benchmark						Khiops		Initial Khiops		CHAID		Tschuprow		Gain Ratio	
Test	Taille	Att. Desc.		Att. Cible		Group Nb.	DKL	Group Nb.	DKL	Group Nb.	DKL	Group Nb.	DKL	Group Nb.	DKL
		Nb	Rep.	Nb.	Rep.										
A1	100	2	1	2	1	1,05	0,66	2,00	1,00	1,04	0,64	2,00	1,00	2,00	1,00
A2	100	3	1	2	1	1,06	0,49	2,00	0,96	1,12	0,59	2,00	0,96	2,00	0,96
A3	1000	3	1	2	1	1,06	0,46	2,00	0,94	1,12	0,56	2,00	0,94	2,00	0,94
A4	1000	5	1	2	1	1,08	0,33	2,01	0,86	1,32	0,58	2,07	0,87	2,00	0,84
A5	1000	10	1	2	1	1,07	0,20	2,15	0,82	1,85	0,73	2,14	0,81	2,15	0,74
A6	1000	50	1	2	1	1,06	0,08	1,49	0,03	3,52	1,02	2,17	0,77	3,29	0,65
A7	10000	3	1	2	1	1,06	0,44	2,00	0,94	1,13	0,55	2,00	0,94	2,00	0,94
A8	10000	5	1	2	1	1,08	0,34	2,01	0,86	1,32	0,58	2,05	0,87	2,00	0,83
A9	10000	10	1	2	1	1,07	0,18	2,18	0,81	1,85	0,71	2,19	0,79	2,12	0,72
A10	10000	50	1	2	1	1,05	0,07	3,63	0,90	3,42	0,88	2,19	0,69	3,19	0,55
A11	10000	100	1	2	1	1,05	0,05	3,86	0,44	4,57	0,94	2,18	0,68	4,61	0,53
A12	10000	500	1	2	1	1,03	0,02	1,00	0,00	8,67	1,19	2,16	0,73	28,78	0,61
A13	10000	50	1	5	1	1,06	0,04	7,20	0,68	6,41	0,65	5,28	0,58	10,93	0,53
A14	10000	50	1	10	1	1,07	0,03	8,82	0,55	7,89	0,52	10,52	0,60	16,55	0,53
A15	10000	50	2	5	1	1,07	0,04	7,18	0,68	6,43	0,65	5,28	0,57	11,76	0,52
A16	10000	50	5	5	1	1,05	0,03	6,81	0,59	6,50	0,64	5,57	0,58	14,51	0,52
A17	10000	50	10	5	1	1,02	0,03	6,13	0,48	6,56	0,63	5,81	0,58	16,85	0,51
A18	10000	50	1	5	2	1,06	0,04	7,10	0,68	6,43	0,65	5,20	0,57	10,91	0,53
A19	10000	50	1	5	5	1,05	0,03	7,12	0,68	6,47	0,65	5,28	0,58	10,93	0,53
A20	10000	50	1	5	10	1,06	0,04	7,16	0,68	6,44	0,66	5,26	0,58	11,03	0,53
Synthèse						1,06	0,10	4,19	0,46	4,20	0,68	3,67	0,72	7,98	0,65

Le premier enseignement de ces résultats est l'intérêt de la distance de Kullback-Leibler pour estimer la qualité d'un groupage. La mesure du taux d'erreur (non présentée ici) donne elle exactement le même résultat (parfait) quelle que soit la méthode de groupage utilisée, y compris pour les méthodes extrêmes qui soit groupent tout, soit ne groupent rien. 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 (normalisée) très souvent inférieure à 1, 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 valeurs cibles.

L'analyse des résultats montre que la méthode Khiops se comporte ici de façon conforme à ses objectifs, à savoir qu'elle aboutit à 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. Toutes les autres méthodes sur-apprennent en élaborant

systématiquement plusieurs groupes là où il n'y aucune information. Les méthodes Khiops initiale, Tschuprow et Gain Ratio sont contraintes de produire au moins deux groupes. La méthode Khiops initiale obtient des performances légèrement dégradées comparativement à la méthode CHAID, excepté pour les jeux d'essai A6 et A11 et A12 pour lesquels elle est artificiellement avantagée en procédant à des regroupements préalables pour atteindre sa contrainte d'effectif minimum par groupe qui est ici plus importante que la taille moyenne des modalités descriptives originelles. La méthode CHAID se comporte honorablement quand le nombre de valeurs descriptives est faible (typiquement inférieur à une dizaine), mais ses performances se dégradent rapidement quand le nombre de valeurs descriptives devient important. La méthode Tschuprow se comporte de façon très prédictible conformément au biais lié à son critère, en produisant systématiquement un nombre de groupes environ égal au nombre de valeurs cibles. La méthode Gain Ratio produit presque toujours un nombre minimal de groupes (à savoir deux) dans le cas où il n'y a que deux modalités cibles, et n'augmente significativement le nombre de groupes que dans le cas où le nombre de valeurs descriptives devient très important. Par contre, dans le cas où il y a cinq modalités cibles, la méthode Gain Ratio produit un nombre de groupes supérieur à une dizaine et présente un comportement instable en fonction de la nature du jeu d'essai.

En résumé, la méthode Khiops est la seule parmi les méthodes testées qui produise un seul groupe dans le cas d'un attribut descriptif indépendant de l'attribut cible. L'expérimentation montre que la distance de Kullback-Leibler constitue un excellent indicateur de la qualité des groupages.

4.4 Jeux d'essai théoriques

On a utilisé le même type de paramètres que pour les jeux d'essai précédents, en ajoutant un paramètre permettant de contrôler le taux d'indépendance. Ce taux d'indépendance varie entre 0 (dépendance complète) et 1 (indépendance complète). Dans le cas de dépendance complète, le choix d'une modalité descriptive conditionne le choix de la modalité cible selon le protocole suivant. Un seul tirage aléatoire détermine la position des modalités descriptive et cible dans leur distribution respective. Par exemple, dans un cas de cinquante modalités descriptives pour cinq modalités cibles, les dix premières modalités descriptives sont associées à la première modalité cible, les dix suivantes à la deuxième, et ainsi de suite. Dans le cas de trois modalités descriptives pour deux modalités cible, la première modalité descriptive est associée à la première modalité cible, la deuxième est équidistribuée sur les deux modalités cibles, et la troisième modalité descriptive est associée à la seconde modalité cible. Pour les cas intermédiaires entre indépendance totale et dépendance totale, une partie des instances (contrôlée par le taux d'indépendance) suit le schéma d'indépendance et l'autre partie suit le schéma de dépendance. Ainsi, le taux d'indépendance de 80% utilisé dans la plupart des jeux d'essai signifie que dans 80% des cas, la modalité cible est déterminée au hasard indépendamment de la modalité descriptive. Le cas du jeu de test B1 par exemple signifie que la première modalité descriptive a une distribution de probabilités de modalités cibles de 60%-40%, alors que la seconde modalité descriptive à une distribution de 40%-60%.

L'expérimentation est menée 1000 fois sur des échantillons générés aléatoirement pour chaque type de jeu d'essai. La taille et la qualité des groupages sont évaluées comme dans l'expérimentation précédente. Le tableau 3 présente l'ensemble des résultats des expérimentations obtenues avec les méthodes testées.

Cette deuxième expérimentation confirme l'intérêt de la distance 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 connu de façon exacte dans le cas des jeux d'essai théorique de cette expérimentation. Cela justifie son utilisation dans le cas de jeux d'essai réels, pour lesquels on ne connaît pas le groupage optimal.

L'analyse des résultats montre que la méthode Khiops obtient les meilleurs résultats de l'expérimentation, en produisant des groupages de plus petite taille et de meilleure qualité que toutes les autres méthodes. Khiops parvient très souvent à déterminer le nombre optimal de groupes (qui dans le cas des jeux d'essais théoriques utilisés coïncide avec le nombre de valeurs cibles dès que le nombre de valeurs descriptives en est un multiple). Dans les cas très bruités B25 et B26, le taux de bruit devient tel que Khiops réduit le nombre de groupes produits vers 1, car les différences entre les groupes deviennent indiscernables du bruit. Dans les cas plus complexes (B2, B3, B4, B7, B8) où il y a 3 groupes pour deux modalités cibles, distribués selon 60%-40%, 50%-50% et 40%-60%, il faut attendre un effectif suffisant pour identifier systématiquement les trois groupes, qui sinon sont masqués par la forte variance des échantillons de petite taille. Les cas B27 et B28 (cinq groupes pour 3 modalités cibles) sont également bien reconnus, mais les cas B28 et B30 (7 groupes pour 4 modalités cibles) ne produisent en moyenne que 5 groupes, à cause de la diminution des effectifs par modalité descriptive. La méthode Khiops initiale obtient de bons résultats, mais est sensible au bruit en produisant trop de groupes dès que le taux de bruit augmente (B24, B25, B26) ou que le nombre de valeurs descriptives se multiplie (B10, B11 par exemple). Sa contrainte d'effectif minimum par groupe lui permet néanmoins de résister honorablement au sur-apprentissage, mais l'empêche de traiter les cas où les valeurs descriptives sont trop nombreuses (B6, B12). La méthode CHAID obtient des résultats dégradés par rapport à la méthode Khiops initiale, en produisant des groupes de taille plus grande et de qualité moins bonne. Il est à noter que la plupart des jeux d'essai ont des effectifs suffisamment importants par valeur descriptive, et que la contrainte d'effectif minimum de Khiops initiale est alors inactive. Comparer les méthodes Khiops initiale et CHAID revient alors à comparer deux algorithmes basés sur le Khi2, le premier utilisant le critère d'évaluation globalement sur l'ensemble des groupes et le second l'appliquant localement sur chaque paire de groupe. Il ressort de cette évaluation que l'évaluation globale de Khiops initiale produit de meilleurs groupages que l'évaluation locale de CHAID. La méthode Tschuprow paraît obtenir d'excellents résultats sur la plupart des jeux d'essai en produisant le nombre optimal de groupe de façon quasi-parfaite (avec une variance presque nulle sur 1000 expérimentations). On constate en fait que

la méthode Tschuprow produit ici systématiquement le même nombre de groupes que de valeurs cibles, ce qui rend ses performances optimale dans les cas où ce biais de la méthode coïncide avec le biais des jeux d'essai, mais la dégrade dans tous les autres cas (par exemple B7, B8, B27 à B30), cette dégradation entraînant une détérioration extrême de la qualité des partitions (voir B7, B8). Cette méthode se révèle ici incapable de trouver le bon nombre de groupes autrement que par hasard. La méthode Gain Ratio obtient de mauvaises performances, exceptés dans les cas où les modalités descriptives sont facilement groupables (B21, B22, B23). Même quand elle produit le bon nombre de groupes, ces groupes ne sont pas toujours constitués correctement, ce qui dégrade leur qualité (par exemple, dans le cas B5, la DKL est de 1,01 pour 2 groupes au lieu de 0,70 pour les méthodes Khiops et Tschuprow produisant également 2 groupes). Cela provient du critère Gain Ratio qui en cherchant un compromis entre le gain d'entropie et l'entropie des groupes sacrifie parfois la qualité informationnelle en préférant une fusion de deux groupes ayant des distributions différentes (perte informationnelle) mais des effectifs importants (diminution importante de l'entropie des groupes). Dans le cas où il y a deux modalités cible, cette méthode paraît biaisée vers des groupages réduits à deux groupes, ce qui entraîne les mêmes défauts que ceux de la méthode Tschuprow. Dans les cas où il y a plus de 2 modalités cibles, la méthode Gain Ratio produit des groupages à la fois surnuméraires et de mauvaise qualité.

En conclusion, la méthode de groupage Khiops robuste produit des groupages de très bonne qualité, souvent optimaux, sur un domaine très étendu.

Tableau 3 : Nombre de groupes et distance de Kullback-Leibler des méthodes de groupage testées sur trente familles de jeux d'essai dans le cas de dépendance partielle entre attribut descriptif et attribut cible

Benchmark							Khiops		Initial Khiops		CHAID		Tschuprow		Gain Ratio	
Test	Taille	Att. Desc.		Att. Cible		Ind.	Group	DKL	Group	DKL	Group	DKL	Group	DKL	Group	DKL
		Nb	Rep.	Nb.	Rep.	Ratio	Nb.		Nb.		Nb.		Nb.		Nb.	
B1	100	2	1	2	1	0,8	1,57	1,62	2,00	1,00	1,53	1,71	2,00	1,00	2,00	1,00
B2	100	3	1	2	1	0,8	1,30	1,33	2,01	1,06	1,45	1,28	2,00	1,06	2,00	1,07
B3	1000	3	1	2	1	0,8	2,29	2,32	2,52	1,80	2,49	1,86	2,00	2,99	2,00	2,99
B4	1000	5	1	2	1	0,8	2,11	1,59	2,45	1,33	2,46	1,33	2,01	1,62	2,00	2,03
B5	1000	10	1	2	1	0,8	2,01	0,70	2,48	0,77	2,54	0,80	2,01	0,70	2,02	1,01
B6	1000	50	1	2	1	0,8	1,90	0,92	1,49	0,89	4,16	1,11	2,07	0,93	2,96	1,16
B7	10000	3	1	2	1	0,8	3,00	1,00	3,00	1,00	3,00	1,00	2,00	23,83	2,00	23,83
B8	10000	5	1	2	1	0,8	3,02	0,62	3,04	0,65	3,10	0,70	2,00	11,25	2,00	11,25
B9	10000	10	1	2	1	0,8	2,01	0,20	2,21	0,36	2,68	0,59	2,00	0,19	2,00	0,19
B10	10000	50	1	2	1	0,8	2,00	0,20	4,63	0,82	5,23	0,86	2,00	0,20	2,00	0,19
B11	10000	100	1	2	1	0,8	2,03	0,59	5,25	2,45	6,56	0,99	2,00	0,60	2,01	0,74
B12	10000	500	1	2	1	0,8	2,00	0,90	1,00	0,88	10,22	1,22	2,10	0,91	33,74	1,23
B13	10000	50	1	5	1	0,8	5,00	0,12	5,79	0,18	10,44	0,49	5,00	0,12	4,94	0,26
B14	10000	50	1	10	1	0,8	10,00	0,21	10,07	0,22	13,64	0,38	10,00	0,21	9,99	0,22
B15	10000	50	2	5	1	0,8	5,02	0,30	7,02	0,36	11,48	0,58	5,00	0,30	4,88	0,56
B16	10000	50	5	5	1	0,8	5,03	0,29	6,63	0,32	11,32	0,60	5,00	0,29	4,79	0,87
B17	10000	50	10	5	1	0,8	5,06	0,41	6,79	0,35	11,74	0,65	5,00	0,42	4,73	1,12
B18	10000	50	1	5	2	0,8	5,01	0,19	6,31	0,28	10,80	0,54	5,00	0,19	4,73	0,72
B19	10000	50	1	5	5	0,8	5,02	0,24	7,12	0,33	11,12	0,57	5,00	0,25	4,27	1,85
B20	10000	50	1	5	10	0,8	5,09	0,33	8,10	0,40	11,77	0,62	5,00	0,34	4,48	2,43
B21	10000	50	1	5	1	0,5	5,00	0,10	5,07	0,11	10,45	0,49	5,00	0,10	5,00	0,10
B22	10000	50	1	5	1	0,6	5,00	0,10	5,16	0,12	10,40	0,49	5,00	0,10	5,00	0,10
B23	10000	50	1	5	1	0,7	5,00	0,10	5,32	0,14	10,37	0,48	5,00	0,10	5,00	0,11
B24	10000	50	1	5	1	0,9	4,95	0,54	8,24	0,62	10,09	0,73	4,97	0,54	5,53	1,08
B25	10000	50	1	5	1	0,95	1,97	0,60	8,52	0,81	8,31	0,83	5,05	0,71	10,59	0,81
B26	10000	50	1	5	1	0,99	1,07	0,06	7,28	0,69	6,51	0,66	5,26	0,58	10,89	0,54
B27	10000	5	1	3	1	0,8	4,96	1,07	4,98	1,05	5,00	1,01	3,00	5,72	2,64	12,82
B28	10000	8	1	3	1	0,8	4,81	0,88	4,95	0,77	5,12	0,75	3,00	3,08	2,81	6,16
B28	10000	15	1	4	1	0,8	5,05	0,83	5,99	0,69	6,96	0,69	4,00	1,52	3,83	2,47
B30	10000	19	1	4	1	0,8	4,63	0,76	5,86	0,58	7,19	0,65	4,00	1,06	3,87	2,47
Synthèse							3,76	0,43	5,04	0,54	7,27	0,76	3,78	0,64	5,16	1,01

4.5 Benchmarks UCI

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 donnée initial

complet. Les 12 jeux de données, totalisant 230 attributs à grouper, sont résumés dans le tableau 4, dont la dernière colonne présente le taux de bonne prédiction du prédicteur majoritaire.

Tableau 4 : 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

L'expérimentation consiste à grouper chaque attribut et à mesurer le nombre de groupes résultant, ainsi que la distance de Kullback-Leibler entre l'estimation des probabilités cibles avant et après groupage. 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. Afin de déterminer si les différences de résultats entre méthodes sont significatives, les tests de Student de comparaisons par paires ont été évalués au seuil de confiance de 5%.

4.5.1 Qualité des groupages

L'utilisation de la distance de Kullback-Leibler permet d'évaluer très finement la qualité prédictive des méthodes de groupage, en évaluant la probabilité prédite de chaque classe cible. La distance de Kullback-Leibler est très sensible et par exemple s'améliore notablement quand la taille de l'ensemble d'apprentissage augmente et permet ainsi une meilleure estimation des probabilités. Les résultats sont alors synthétisés par des moyennes géométriques plutôt que par des moyennes arithmétiques, et en normalisant la distance par rapport à la distance de référence obtenue dans le cas sans groupage.

La table de résultats complète pour les 230 attributs est trop volumineuse pour être reproduite dans ce document. Elle est résumée dans le tableau 5, qui indique pour chaque jeu de données la moyenne géométrique normalisée des distances de Kullback-Leibler par attribut et le nombre de dégradations (-) et d'améliorations (+) significatives pour Khiops dans les comparaisons avec les autres méthodes.

Tableau 5 : Moyenne par jeu de données des qualités des groupages, nombres de dégradations et d'améliorations significatives pour Khiops

Dataset	Khiops	Ini. Khiops			CHAID			Tschuprow			Gain Ratio		
			-	+		-	+		-	+		-	+
Adult	1,05	1,13	2	3	1,07	4	4	3,76	0	10	4,16	0	10
Australian	1,04	1,06	0	0	1,10	0	2	1,10	0	1	1,24	0	3
Breast	1,24	1,24	0	1	1,36	0	4	1,45	0	2	1,66	0	5
Crx	1,06	1,07	1	0	1,08	1	0	1,10	0	1	1,23	0	3
Heart	0,98	1,02	0	1	1,02	0	0	1,03	0	2	1,07	0	3
HorseColic	1,02	1,01	4	1	1,07	2	3	1,08	0	3	1,04	2	3
Ionosphere	1,07	1,03	3	1	1,13	1	7	1,06	2	2	1,08	4	3
Mushroom	1,10	1,24	2	4	1,21	2	6	2,29	1	11	2,60	1	11
TicTacToe	0,97	0,97	0	0	0,91	1	0	0,95	0	0	0,95	0	0
Vehicle	1,10	1,10	1	5	1,11	4	4	1,12	2	2	1,30	0	9
Waveform	0,92	0,99	0	13	1,01	0	19	1,48	0	30	1,47	0	30
Wine	1,23	1,20	1	0	1,37	1	6	1,24	1	1	1,23	1	0
Synthesis	1,04	1,07	14	29	1,10	16	55	1,35	6	65	1,42	8	80

La valeur de la distance de Kullback-Leibler normalisé est ici en moyenne supérieure à 1, ce qui signifie qu'il y a une dégradation de l'estimation des probabilités cible après groupage par rapport à leur estimation avant groupage à partir des valeurs descriptives originelles. Dans ces jeux d'essai "réels", le meilleur groupage est un compromis entre dégradation

minimale de la qualité prédictive et diminution maximale du nombre de valeurs. Si l'on groupe deux valeurs descriptives ayant la même distribution de valeurs cible, on améliore la qualité de l'estimation qui est basée sur un effectif plus important (ce qui était le cas dans les jeux d'essai théoriques). Si par contre on regroupe à tort deux valeurs descriptives dissemblables, on détériore la qualité du groupage.

Les différences de résultats entre méthodes sont significatives et permettent d'ordonner complètement les méthodes. Dans un premier groupe assez resserré, la méthode Khiops obtient les meilleures performances, suivi des méthodes Khiops initiale puis CHAID. Khiops est meilleure que CHAID significativement dans 24% des cas de groupage, et n'est surclassée que dans 7% des cas. Les méthodes Tschuprow puis Gain Ratio arrivent très loin derrière les trois meilleures méthodes. Ainsi; Khiops est meilleure que Gain Ratio significativement dans 35% des cas de groupage, et n'est surclassée que dans 3,5% des cas. Les deux méthodes Tschuprow et Gain Ratio ont des performances d'autant plus dégradées que les tailles des échantillons sont importantes, comme on peut le voir pour les jeux d'essai Adult, Mushroom et Waveform.

L'approche d'évaluation globale des groupes de Khiops initiale permet d'obtenir de meilleurs résultats que l'approche d'évaluation locale par paires de groupes de CHAID. La méthode Khiops améliore significativement la méthode Khiops initiale en apportant une garantie statistique sur la qualité des groupages produits. De plus, le relâchement de la contrainte d'effectifs minimums lui permet d'identifier des groupes de plus faible taille. La méthode Tschuprow souffre de son biais qui privilégie très fortement des partitions ayant le même nombre de groupes que de modalités cibles, ce qui l'empêche de trouver les bons groupages. La méthode Gain Ratio pâtit essentiellement de son critère heuristique qui tente de concilier information prédictive (le gain d'entropie) et compacité du groupage (entropie des groupes) en en faisant un ratio. Ce critère heuristique conduit alors à une optimisation soit du numérateur, soit du dénominateur, et rarement à un compromis équilibré entre les deux facteurs.

4.5.2 Taille des groupages

La taille des groupages est simplement le nombre de groupes obtenus sur l'ensemble d'apprentissage par les différentes méthodes testées. Le tableau 6 résume les résultats obtenus lors de l'évaluation de la taille des groupages sur l'ensemble des attributs des jeux de données.

Tableau 6 : Moyenne par jeu de données des tailles des groupages, nombres de dégradations et d'améliorations significatives pour Khiops

Dataset	Khiops	Ini. Khiops			CHAID			Tschuprow			Gain Ratio		
			-	+		-	+		-	+		-	+
Adult	3,67	3,99	0	5	4,83	0	11	2,05	10	2	2,33	10	2
Australian	1,91	2,19	1	6	2,19	0	4	2,19	1	3	2,36	1	7
Breast	2,60	2,83	0	3	4,16	0	9	1,98	7	1	1,98	7	1
Crx	1,93	2,16	1	5	2,18	0	3	2,15	1	3	2,42	2	8
Heart	1,91	2,27	0	5	2,14	0	4	2,11	1	3	2,08	1	3
HorseColic	1,87	2,20	0	11	2,24	0	10	2,03	4	7	2,03	4	8
Ionosphere	2,47	2,94	1	17	3,18	0	25	2,09	15	0	2,05	17	0
Mushroom	3,06	3,11	3	3	3,57	0	10	2,00	13	0	2,19	13	1
TicTacToe	2,03	2,03	0	0	2,11	0	1	2,00	0	0	2,00	0	0
Vehicle	3,50	3,90	0	7	4,84	0	17	2,58	11	0	2,85	11	3
Waveform	2,67	3,56	0	30	3,76	0	35	2,73	19	21	3,18	18	21
Wine	2,60	2,95	0	5	3,56	0	11	2,10	6	0	2,05	7	1
Synthesis	2,54	2,95	6	97	3,28	0	140	2,22	88	40	2,38	91	55

L'analyse des résultats permet de retrouver les deux ensembles de méthodes précédents, avec d'une part les méthodes performantes Khiops, Khiops initiale et CHAID qui produisent au moins autant de groupes que la méthode Khiops, et d'autre part les méthodes peu performantes qui produisent moins de groupes que la méthode Khiops, au détriment d'une perte très importante de la qualité de groupage. Parmi les méthodes performantes, la méthode Khiops produit les groupages les plus compacts de façon très significative. Ainsi, les groupages de Khiops sont toujours plus compacts que ceux de CHAID, et ce de façon significative dans plus de 60% des cas.

Le cas du jeu d'essai Waveform est très illustratif du comportement des méthodes. Ce jeu d'essai élaboré par (Breiman 1984) comporte pour moitié des attributs informatifs et pour moitié des attributs de bruit. La méthode Khiops obtient les groupages de meilleure qualité et les plus compacts. Les attributs informatifs donnent lieu à 4 ou 5 groupes en moyenne, tandis que les attributs de bruit aboutissent à un seul groupe. Les méthodes Khiops initiale et CHAID sur-apprennent en produisant des groupes en moyenne 30% plus nombreux que ceux de Khiops, ce qui dégrade légèrement leur qualité prédictive, de façon assez uniforme sur l'ensemble des attributs informatifs et de bruit. Les méthodes Tschuprow et Gain Ratio produisent exactement deux groupes pour les attributs informatifs, ce qui n'est pas assez, et plus de trois en moyenne (pour Tschuprow) ou quatre (pour Gain ratio) dans le cas des attributs de bruit, ce qui est trop. Dans tous les cas, la qualité des groupages est alors sévèrement dégradée.

4.5.3 Synthèse des résultats

On va ici procéder à une analyse bi-critères de la qualité et de la taille des groupages, en intégrant les résultats de méthodes produisant des bipartitions ou des partitions élémentaires. Ces méthodes sont:

- Exhaustive CHAID: variante de l'algorithme CHAID forçant les groupages jusqu'à l'obtention d'au plus deux groupes
- Chi Single Value: choix d'une modalité contre toutes les autres, en se basant sur le critère du Khi2 au seuil de 95% (ce qui engendre soit une bipartition, soit un seul groupe terminal)
- Mode: choix de la modalité la plus fréquente contre toutes les autres
- One Group: fusion de toutes modalités en un seul groupe
- No Grouping: aucun groupage n'est effectué

L'évaluation de toutes ces méthodes permet de les situer par rapport au compromis entre compacité et qualité des groupages. Le tableau 7 présente les résultats des groupages de 230 attributs de façon synthétique, en reportant la moyenne des distances de Kullback-Leibler et des tailles de groupages, ainsi que le nombre de différences significatives dans les comparaisons avec Khiops.

Tableau 7 : Moyenne globale des qualités et tailles des groupages, nombre de dégradations et d'améliorations significatives pour Khiops. Evaluation des méthodes de groupages multi-groupes, bi-groupes et élémentaires

Grouping Method	Kullback-Leibler Distance	Khiops losses	Khiops wins	Group Number	Khiops losses	Khiops wins
Khiops	1,04			2,54		
Initial Khiops	1,07	14	29	2,95	6	97
CHAID	1,10	16	55	3,28	0	140
Tschuprow	1,35	6	65	2,22	88	40
Gain Ratio	1,42	8	80	2,38	91	55
Exhaustive CHAID	1,36	5	66	1,90	98	23
Chi Single Value	1,54	18	64	1,86	98	12
Mode	1,74	20	63	1,99	98	40
One group	2,69	12	101	1,00	190	0
No Grouping	1,00	47	30	7,89	0	213

La valeur de la distance de Kullback-Leibler pour les méthodes élémentaires permet d'étalonner l'ensemble des résultats et d'évaluer les performances relatives entres les méthodes. On remarque que la dégradation de qualité prédictive est minime entre Khiops et le cas où aucun groupage n'est effectué (No Grouping), alors que le nombre de groupes est divisé en moyenne par 3. La figure 1 présente l'ensemble des méthodes évaluées sur les deux facteurs de taille et qualité des groupages.

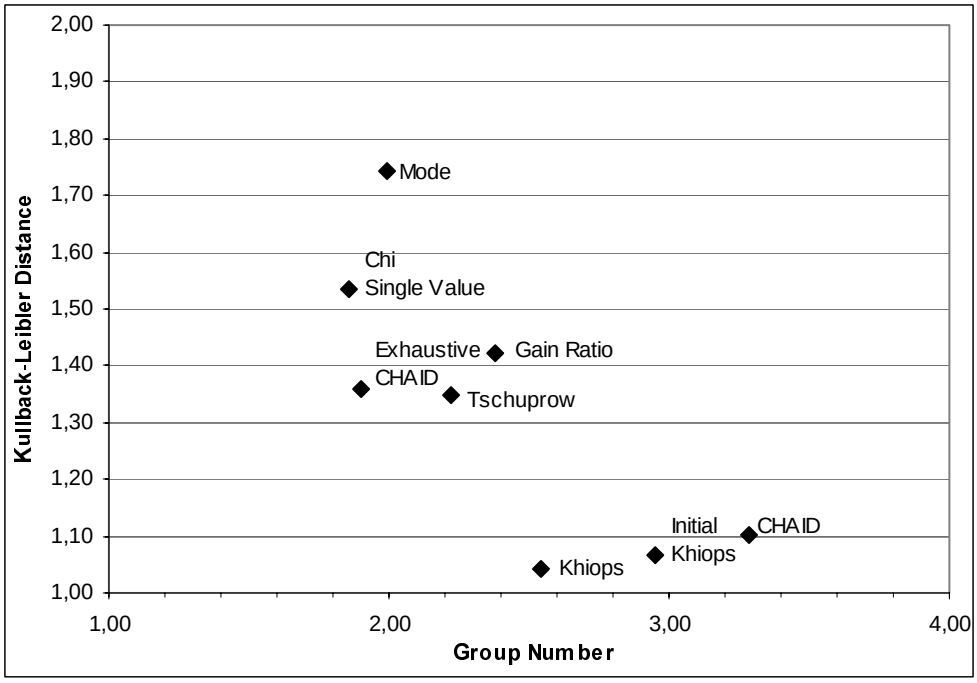

Figure 1 : Taille et qualité des partitions produites par les méthodes de groupage testées sur les bases de l'UCI

La qualité des bipartitions s'améliore quand on passe de la méthode élémentaire Mode qui isole la valeur descriptive la plus fréquente à la méthode Single Value qui isole la valeur descriptive la plus informative, puis quand on utilise la méthode Exhaustive CHAID qui recherche la bipartition optimale. Les deux méthodes Tschuprow et Gain Ratio n'apportent aucune amélioration de qualité par rapport à la méthode Exhaustive CHAID en dépit d'une augmentation importante du nombre moyen de groupes. Les trois méthodes performantes Khiops, Khiops initiale et CHAID se détachent clairement au bas de la figure, Khiops dominant nettement les deux autres méthodes sur les deux critères.

Conclusion

La méthode Khiops groupe les modalités d'un attribut symbolique en minimisant la probabilité d'indépendance entre attribut groupé et attribut cible. Lors d'un groupage, de nombreuses fusions de modalités sont effectuées, donnant lieu à des variations DeltaKhi2 de la valeur du Khi2 du tableau de contingence. Ces variations sont bornée par leur valeur maximale MaxDeltaKhi2 lors du déroulement complet de l'algorithme. Nous avons montré que dans le cas d'un attribut descriptif indépendant d'un attribut cible, la variable MaxDeltaKhi2 ne dépend que du nombre de modalités descriptives et du nombre de modalités cibles, qu'elle est insensible à la taille de l'échantillon et à la répartition des modalités descriptives et cibles. De plus, la valeur MaxDeltaKhi2 est en relation quasi-linéaire avec les nombres de modalités descriptives et cibles, ce qui permet son évaluation par interpolation à partir d'une table de valeurs calculée expérimentalement. Cette connaissance de la statistique du MaxDeltaKhi2 nous a permis d'améliorer la robustesse de l'algorithme de groupage en imposant à l'algorithme d'accepter toute fusion entraînant une variation du Khi2 inférieure à ce MaxDeltaKhi2. La version robuste de l'algorithme Khiops apporte alors la garantie que les attributs sans intérêt prédictif sont groupés en un seul groupe terminal. Des expérimentations ont permis de valider ces analyses, puis montré que la méthode Khiops robuste conduit à des résultats de grande qualité dans de très larges gammes de types de jeux d'essai. Cette approche permet de contrôler le problème de sur-apprentissage a priori, et constitue une alternative intéressante à l'approche classique de contrôle du sur-apprentissage a posteriori par utilisation d'échantillons de validation.

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. (2002). Amélioration de la robustesse de la méthode de discrétisation Khiops par contrôle de son comportement statistique. Note technique NT/FTR&D/7864; 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.

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.

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.

5 Annexe : étude de la statistique du MaxDeltaKhi2 de groupage

5.1 Présentation

Le MaxDeltaKhi2 est la valeur maximale du DeltaKhi2 observée lors du groupage total (jusqu'à un unique groupe terminal) d'un attribut descriptif indépendant de l'attribut cible. Cette valeur MaxDeltaKhi2 dépend a priori de la taille de l'échantillon, des nombres de modalités descriptives et cibles, et de leur répartition de fréquences. N'étant pas en mesure de modéliser de façon analytique le comportement du MaxDeltaKhi2 dans le cas général, l'étude de la statistique du MaxDeltaKhi2 de groupage est empirique, et ses résultats ne sont pas démontrés, mais vérifiés expérimentalement.

Nous avons utilisé les paramètres suivant pour cette étude empirique:

- N : taille de l'échantillon
- I : nombre de modalités descriptives
- J : nombre de modalités cibles
- DFD : taux de Déséquilibre des Fréquences des modalités Descriptives
- DFC : taux de Déséquilibre des Fréquences des modalités Cibles

Le taux de déséquilibre est égal au rapport entre la fréquence la plus importante et la fréquence la plus faible, l'ensemble des fréquences suivant une loi géométrique.

Pour chaque jeu de paramètres les statistiques sont collectées en effectuant 1000 groupages (jusqu'à l'obtention d'un seul groupe terminal) pour des jeux d'essai générés aléatoirement avec indépendance des attributs descriptif et cible. Ces groupages sont faits en se plaçant dans les conditions où il n'y a pas de problème d'effectif minimum dans les groupes initiaux. Les résultats sont présentés sous forme de la fonction de répartition pour différentes valeurs d'un paramètre étudié, en fixant tous les autres paramètres.

5.2 Insensibilité à la taille de l'échantillon

Le MaxDeltaKhi2 est indépendant de la taille de l'échantillon. Cette assertion est vérifiée expérimentalement en faisant varier la taille de l'échantillon de 1000 jusqu'à 200000, pour des jeux d'essai comportant 50 modalités descriptives équidistribuées et 2 modalités cibles équidistribuées. La figure 2 montre que les fonctions de répartition du MaxDeltaKhi2 sont confondues pour toutes les valeurs de la taille de l'échantillon.

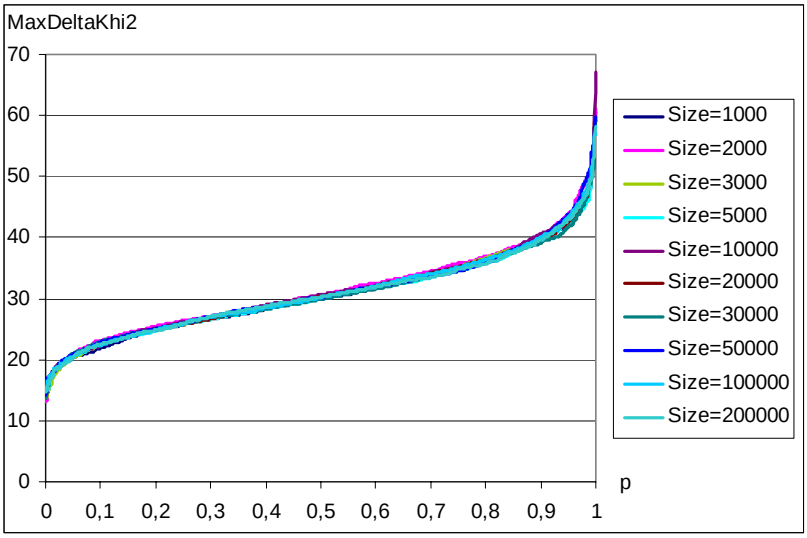

Figure 2 : Insensibilité à la taille de l'échantillon (50 modalités descriptives, 2 modalités cibles)

Ce résultat intéressant découle en fait des propriétés du test du Khi2 qui mesure une probabilité d'indépendance entre deux attributs. Pour deux attributs effectivement indépendants, le test du Khi2 se comporte de façon indépendante de la taille de l'échantillon.

5.3 Insensibilité à la répartition des modalités cibles

Le MaxDeltaKhi2 est indépendant de la répartition des modalités cibles. Cette assertion est vérifiée expérimentalement en faisant varier le ratio de fréquence entre les modalités cibles majoritaire et minoritaire de 1 à 10, pour des jeux d'essai comportant 50 modalités descriptives équidistribuées et 2 modalités cibles. La figure 3 montre que les fonctions de répartition du MaxDeltaKhi2 sont confondues pour toutes les valeurs de ce ratio.

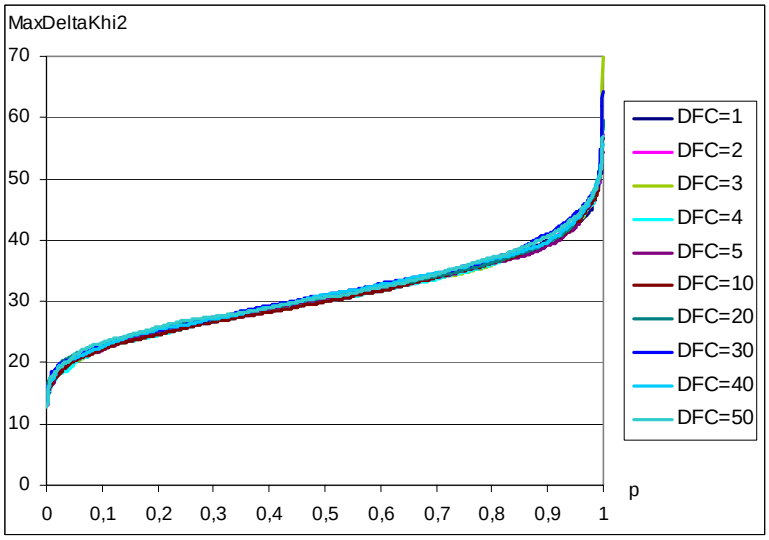

Figure 3 : Insensibilité à la répartition des modalités cibles (50 modalités descriptives, 2 modalités cibles)

Cette propriété découle également du test du Khi2, qui dans sa formule pondère les écarts quadratiques entre probabilités observées et probabilités attendues par l'inverse des probabilités attendues. Cette pondération conduit naturellement à se rendre insensible de la distribution des probabilités cibles, se retrouve dans la valeur moyenne du Khi2 qui ne dépend que du nombre de modalités, indépendamment de leur répartition.

5.4 Insensibilité à la répartition des modalités descriptives

Le MaxDeltaKhi2 est indépendant de la répartition des modalités descriptives. Cette assertion est vérifiée expérimentalement en faisant varier le ratio de fréquence entre les modalités descriptives majoritaire et minoritaire de 1 à 10, pour des jeux d'essai comportant 50 modalités descriptives et 2 modalités cibles équidistribuées. La figure 4 montre que les fonctions de répartition du MaxDeltaKhi2 sont très proches pour l'ensemble des valeurs de ce ratio.

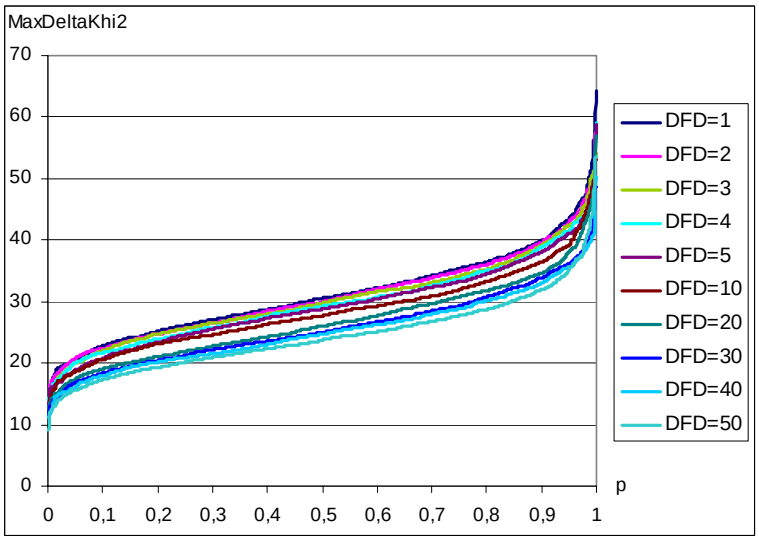

Figure 4 : Quasi-insensibilité à la répartition des modalités descriptives (50 modalités descriptives, 2 modalités cibles)

On peut faire ici la même remarque que pour l'insensibilité à la répartition des modalités cibles. Il y a néanmoins une différence avec le cas précédent dans le sens où les modalités descriptives sont elles soumises à l'algorithme de groupage, qui est une heuristique gloutonne et aura tendance à différencier les regroupements selon les effectifs par modalité descriptive. Cela peut expliquer la quasi-insensibilité observée empiriquement, au lieu de l'insensibilité observée dans les cas des modalités cibles.

On observe que les fonctions de répartition sont ordonnées selon le taux de déséquilibre des modalités cibles. La fonction de répartition correspondant au cas équidistribué est au dessus de toutes les autres et peut raisonnablement être considérée comme une borne sup. En fait, l'algorithme de groupage étant une heuristique, il a d'autant plus de mal à trouver la solution optimale que le nombre de modalités descriptives est important.

5.5 Cas avec deux modalités descriptives

Dans le cas particulier où il n'y a que deux modalités descriptives, il n'y a qu'une seule fusion possible, et le MaxDeltaKhi2 est donc égal au DeltaKhi2, qui suit ici une loi du Khi2 à J-1 degrés de libertés. Donc, pour I=2, MaxDeltaKhi2 vaut J-1 en moyenne.

5.6 Cas avec deux modalités cibles

Dans le cas particulier où il n'y a que deux modalités cibles, on a essayé d'estimer la valeur moyenne du MaxDeltaKhi2 de façon analytique.

Proposition: Dans le cas de I modalités descriptives équidistribuées et 2 modalités cibles équidistribuées, la valeur moyenne du MaxDeltaKhi2 est asymptotiquement linéaire par rapport au nombre de modalités descriptives avec un coefficient de $2/\pi$.

Preuve:

Pour deux modalités descriptives d'effectif n et n', de distribution de probabilités cibles locales p_j et p'_j et pour une distribution globale P_j de modalités cible, on a :

$$DeltaKhi2 = \frac{nn'}{n+n'}\sum_{j=1}^{J}\frac{(p_j - p'_j)^2}{P_j}$$

Si on accepte l'hypothèse que le MaxDeltaKhi2 ne dépend ni de la répartition des modalités descriptives, ni de la répartition des modalités cibles, ni de la taille de l'échantillon, on peut se placer dans le cas où les modalités descriptives ont toutes le même effectif, et où cet effectif est suffisamment grand.

En posant $p=p_1 = 1-p_2$, $p'=p'_1 = 1-p'_2$, $P=P_1 = 1-P_2$ et en se plaçant dans le cas où n=n' :

$$DeltaKhi2 = \frac{n}{2P(1-P)}(p-p')^2$$

Pour P=1/2, le MaxDeltaKhi2 est atteint quand il ne reste que deux groupes terminaux de cardinalité identique N/2, le premier regroupant toutes les modalités descriptives ayant un p <= P, et le second les autres modalités.

Pour n suffisamment grand, p suit une loi normale de moyenne P et de variance P(1-P)/n. Calculons la moyenne et la variance des valeurs de p supérieures à P.

La densité de la loi normale est $\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(t-m)^2}{2\sigma^2}}$.

La moyenne des valeurs supérieures à la moyenne est égale à :

$$M^+ = \frac{1}{\sigma\sqrt{2\pi}}\int_m^\infty te^{-\frac{(t-m)^2}{2\sigma^2}}dt \Bigg/ \frac{1}{\sigma\sqrt{2\pi}}\int_m^\infty e^{-\frac{(t-m)^2}{2\sigma^2}}dt$$

$$M^+ = \frac{2}{\sqrt{2\pi}}\int_0^\infty (\sigma t + m)e^{-\frac{t^2}{2}}dt$$

$$M^+ = 2\sigma\frac{1}{\sqrt{2\pi}}\left[-e^{-\frac{t^2}{2}}\right]_0^\infty + 2m\frac{1}{\sqrt{2\pi}}\int_0^\infty e^{-\frac{t^2}{2}}dt$$

$$M^+ = \frac{2\sigma}{\sqrt{2\pi}} + m$$

$$M^+ = \sqrt{\frac{2P(1-P)}{\pi n}} + P$$

La variance des valeurs supérieures à la moyenne est :

$$V^+ = \frac{2}{\sigma\sqrt{2\pi}}\int_m^\infty (t-M^+)^2 e^{-\frac{(t-m)^2}{2\sigma^2}}dt$$

$$V^+ = \frac{2\sigma^2}{\sqrt{2\pi}}\int_0^\infty \left(t-\sqrt{\frac{2}{\pi}}\right)^2 e^{-\frac{t^2}{2}}dt$$

$$V^+ = 2\sigma^2\left(\frac{1}{\sqrt{2\pi}}\int_0^\infty t^2 e^{-\frac{t^2}{2}}dt - \frac{2}{\pi}\int_0^\infty te^{-\frac{t^2}{2}}dt + \frac{2}{\pi}\frac{1}{\sqrt{2\pi}}\int_0^\infty e^{-\frac{t^2}{2}}dt\right)$$

$$V^+ = \sigma^2\left(1-\frac{2}{\pi}\right)$$

L'écart type est :

$$\sigma^+ = \sigma\sqrt{1-\frac{2}{\pi}}$$

$$\sigma^+ = \sqrt{\frac{P(1-P)}{n}\left(1-\frac{2}{\pi}\right)}$$

Si on construit le premier groupe avec l'ensemble des I^+=I/2 groupes ayant les proportions p les plus grandes, on a :

$$MaxDeltaKhi2 = \frac{nI^+}{2P(1-P)}\left(2P^+ - 1\right)^2$$

$$MaxDeltaKhi2 = \frac{2n}{P(1-P)}\left(\sqrt{I^+}\left(P^+ - P\right)\right)^2$$

$$MaxDeltaKhi2 = \frac{2n}{P(1-P)}\left(V^+\left(\sqrt{I^+}\,\frac{P^+ - M^+}{\sigma^+}\right)^2 + 2\sqrt{I^+}\sigma^+\left(\sqrt{I^+}\,\frac{P^+ - M^+}{\sigma^+}\right)\left(M^+ - P\right) + I^+\left(M^+ - P\right)^2\right)$$

Le terme $\sqrt{I^+}\,\frac{P^+ - M^+}{\sigma^+}$ converge vers la loi normale, et son carré vers la loi du Khi2 à un degré de liberté.

Donc, on obtient la valeur moyenne suivante :

$$MaxDeltaKhi2 = \frac{2n}{P(1-P)}\left(V^+ + I^+\left(M^+ - P\right)^2\right)$$

$$MaxDeltaKhi2 = \frac{2n}{P(1-P)}\left(\frac{P(1-P)}{n}\left(1-\frac{2}{\pi}\right) + I^+\,\frac{2P(1-P)}{\pi n}\right)$$

$$MaxDeltaKhi2 = 2\left(1-\frac{2}{\pi}\right) + \frac{2}{\pi}I$$

Donc dans le cas où P=1/2, où les modalités descriptives sont équidistribuées et où il y a deux modalités cibles équidistribuées, la valeur MaxDeltaKhi2 est asymptotiquement linéaire par rapport au nombre de modalités descriptives avec un coefficient de $2/\pi$.

5.7 Cas général

Dans le cas général, on conjecture que la valeur MaxDeltaKhi2 ne dépend que des nombres de modalités descriptives et cibles. De plus, en se basant sur le comportement théorique obtenu pour I=2 et J=2, on conjecture également que cette dépendance est approximativement linéaire dès que les nombres de modalités sont suffisamment grands.

Cette dernière conjecture est confirmée par les expérimentations présentées sur la figure 5.

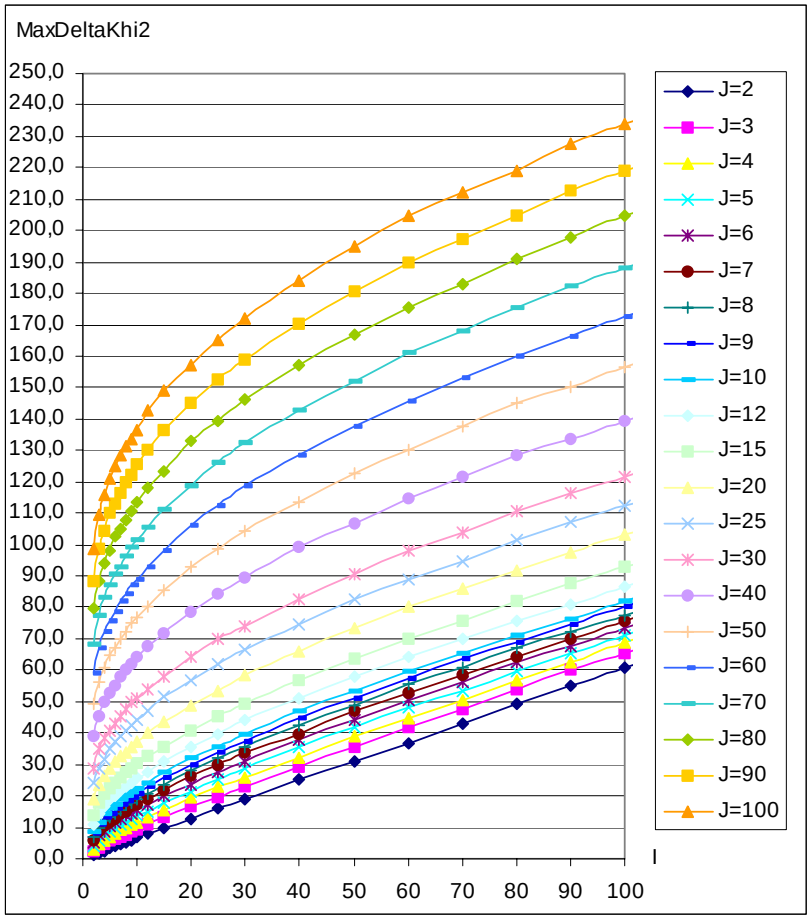

Figure 5 : Moyenne du MaxDeltaKhi2 pour différents (I,J), calculée sur 1000 groupages

5.8 Approximation de la loi de MaxDeltaKhi2 par la loi normale

On conjecture que l'on peut approximer la fonction de répartition du MaxDeltaKhi2 par la loi normale de même moyenne et écart type. Cette conjecture est confirmée par les expérimentations présentées sur la figure 6, sur laquelle on a tracé la fonction de répartition du MaxDeltaKhi2 résultant de 1000 groupages pour 50 modalités descriptives équidistribuées et 2 modalités cibles équidistribuées.

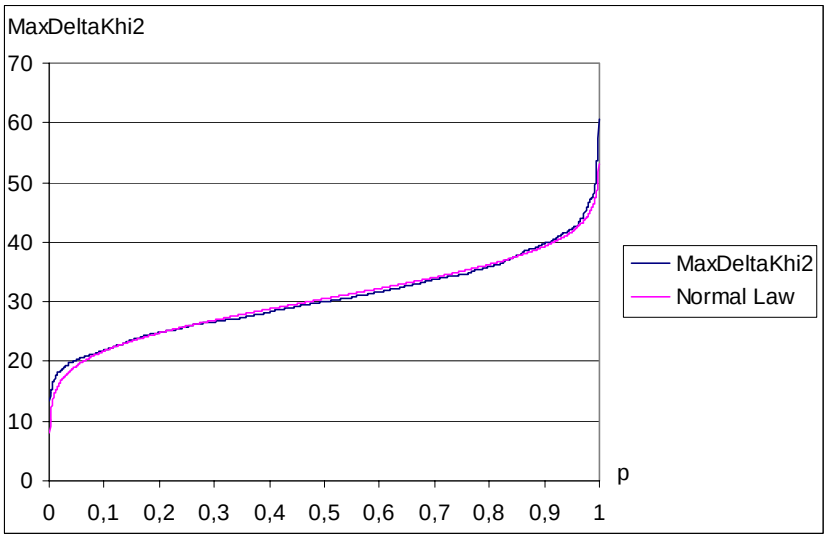

Figure 6 : Fonction de répartition du MaxDeltaKhi2 pour I = 50 et J = 2, et son approximation par la loi normale

5.9 Remarque

Si l'on compare le comportement du MaxDeltaKhi2 pour la discrétisation et le groupage, on observe des différences extrêmement importantes. Dans le cas de la discrétisation, le problème est très fortement contraint par la relation d'adjacence entre intervalles, ce qui a pour conséquence que l'ensemble des DeltaKhi2 de toutes les fusions d'intervalles lors d'une exécution complète de l'algorithme suit approximativement la même loi que la loi de fusion de deux intervalles quelconques. Tout se passe comme si les fusions d'intervalles d'un attribut continu étaient quasiment indépendantes les unes des autres sur l'ensemble du domaine numérique discrétisé (pour un attribut descriptif indépendant de l'attribut cible)

Dans le cas du groupage, toutes les fusions possibles sont envisagées. Ainsi, chaque fusion a une influence directe sur toutes les autres. La loi du MaxDeltaKhi2 est alors radicalement différente. Pour mémoire, dans le cas de deux modalités cibles, le DeltaKhi2 suit une loi du Khi2 à un degré de liberté, et a une valeur moyenne de 1 et une variance de 2. Dans le cas d'une discrétisation d'un échantillon de taille N, le MaxDeltaKhi2 suit une loi correspondant à la plus grande valeur du DeltaKhi2 parmi N possibilités. Cela conduit à un MaxDeltaKhi2 valant environ 7 pour N=100 et 20 pour N=1000000 (donc 20 fois la valeur moyenne). Dans le cas d'un groupage de I modalités descriptives, la valeur moyenne du MaxDeltaKhi2 est d'environ $2I/\pi$. C'est à dire que pour N=500, on peut avoir au moins une trentaine de modalités (respectant l'effectif minimum) et obtenir un MaxDeltaKhi2 dépassant 20 (valeur atteinte pour une discrétisation avec N=1000000). Pour N=1000000, on peut alors obtenir un MaxDeltaKhi2 plusieurs dizaines de milliers de fois plus important que la valeur moyenne du DeltaKhi2.