NT/FTR&D/7864

Janvier 2003

Amélioration de la robustesse de la méthode Khiops par contrôle de son comportement statistique.

DTL
Direction des
Techniques
Logicielles

Marc Boullé (DTL/TIC)

Note Technique

diffusion libre

france telecom
R&D

Note Technique
NT/FTR&D/7864

6 janvier 2003

Amélioration de la robustesse de la méthode Khiops par contrôle de son comportement statistique

Marc Boullé (DTL/TIC)

Vu, pour accord le directeur de DTL

J.M. Pitié

Vu, le responsable dulaboratoire TIC

O. Perrault

Date : 6 janvier 2003

Résumé : Dans le domaine de l'apprentissage supervisé, les méthodes de discrétisation des attributs continus partitionnent un domaine numérique en un nombre fini d'intervalles. La méthode Khiops optimise le critère du Khi2 globalement sur l'ensemble du domaine de discrétisation. Nous proposons ici une évolution majeure de Khiops basée sur l'analyse statistique du déroulement de son algorithme. Cette évolution permet de contrôler a priori le risque de sur-apprentissage et d'améliorer ainsi significativement la robustesse des discrétisations produites. Des expérimentations intensives ont été menées sur de nombreux jeux de données de nature différente, afin de comparer la méthode Khiops avec d'autres méthodes de discrétisation de référence en prenant en compte plusieurs critères tels la performance prédictive, la robustesse ou la taille des discrétisations. Une analyse multi-critères des résultats démontre les bonnes performances de la méthode Khiops, tout en apportant un éclairage intéressant sur la problématique générale de la discrétisation.

Mots clés : **analyse intelligente donnée; apprentissage automatique; discrétisation**

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 4 615 327 772 € - 380 129 866 RCS Paris

Amélioration de la robustesse de la méthode de discrétisation Khiops par contrôle de son comportement statistique

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 discrétisation des attributs continus partitionnent un domaine numérique en un nombre fini d'intervalles, en recherchant un compromis entre valeur informationnelle et valeur prédictive de la partition formée. La méthode Khiops[*] optimise le critère du Khi2 globalement sur l'ensemble du domaine de discrétisation, alors que les méthodes apparentées ChiMerge et ChiSplit optimisent ce critère localement à deux intervalles adjacents. Nous proposons ici une évolution majeure de l'algorithme de discrétisation Khiops permettant de contrôler a priori le risque de sur-apprentissage et d'améliorer ainsi significativement la robustesse des discrétisations produites. Cette amélioration se base sur l'étude de la statistique des variations de la valeur du Khi2 lors de regroupements de lignes d'un tableau de contingence et sur une modélisation du comportement statistique de l'algorithme Khiops. Cette modélisation, vérifiée expérimentalement, permet d'améliorer l'algorithme de discrétisation en offrant des garanties concernant le sur-apprentissage. Des expérimentations intensives ont été menées sur de nombreux jeux de données de nature différente, afin de comparer la méthode Khiops avec d'autres méthodes de discrétisation de référence en prenant en compte plusieurs critères tels la performance prédictive, la robustesse ou la taille des discrétisations. Une analyse multi-critères des résultats démontre les bonnes performances de la méthode Khiops, tout en apportant un éclairage intéressant sur la problématique générale de la discrétisation.

Mots clés : Data Mining, Machine Learning, Discretization, Data Analysis

---

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

HISTORIQUE

Septembre 2001 NT/FTR&D/7339 : Khiops : « Discrétisation des attributs numériques pour le Data Mining »

- Note technique présentant la première version de la méthode Khiops

Novembre 2002 NT/FTR&D/7864 : « Amélioration de la robustesse de la méthode de discrétisation Khiops par contrôle de son comportement statistique »

- Chapitres 2 et 3 : Simplification de la rédaction des chapitres originels sur la présentation de la méthode Khiops initiale et sur sa comparaison théorique avec ChiMerge et ChiSplit
- Chapitre 4 : Nouveau chapitre introduisant les améliorations de la méthode
- Chapitre 5 : Remplacement du chapitre précédent sur les expérimentations sur des jeux de données théoriques par un chapitre portant sur des expérimentations sur des bases réelles avec une analyse multi-critères des résultats
- Annexe 6 : inchangé par rapport à la version précédente

Principales modifications apportées à la méthode Khiops initiale :

- Le contrôle du sur-apprentissage est désormais garanti de façon théorique (chapitre 4.2), alors qu'il était auparavant géré de façon heuristique en imposant un effectif minimum par intervalle.
- Des expérimentations comparatives intensives ont été menées, donnant lieu à une analyse multi-critères des résultats (chapitre 5).

TABLE DES MATIERES

1 Introduction

La discrétisation des attributs numériques est un sujet largement traité dans la bibliographie (Zighed et Rakotomalala 2000). Une partie des modèles d'apprentissage est basée sur le traitement des attributs à valeurs discrètes. Il est donc nécessaire de discrétiser les attributs numériques, c'est à dire de découper leur domaine en un nombre fini d'intervalles identifiés chacun par un code. Ainsi, tous les modèles prédictifs à base d'arbre de décision utilisent une méthode de discrétisation pour traiter les attributs numériques. C4.5 (Quinlan 1993) utilise le gain informationnel basé sur l'entropie de Shannon, CART (Breiman 1984) utilise l'indice de Gini (une mesure de l'impureté des intervalles), CHAID (Kass 1980) s'appuie sur une méthode de type ChiMerge, SIPINA utilise le critère Fusinter (Zighed 1998) basé sur des mesures d'incertitude sensibles aux effectifs.

Parmi les méthodes de discrétisation, il existe des méthodes descendantes et ascendantes. Les méthodes descendantes partent du domaine numérique complet à discrétiser et le coupent en deux récursivement. Les méthodes ascendantes partent des intervalles élémentaires mono-valeur et les fusionnent itérativement. Certaines de ces méthodes nécessitent un paramétrage utilisateur pour modifier le comportement du critère de choix du point de discrétisation ou pour fixer un seuil pour le critère d'arrêt. Le problème de la discrétisation est un problème de compromis entre qualité informationnelle (intervalles homogènes vis à vis de la variable à prédire) et qualité statistique (effectif suffisant dans chaque intervalle pour assurer une généralisation efficace). Les critères de type Khi2 privilégient l'aspect statistique tandis que ceux basés sur la mesure de l'entropie privilégient l'aspect informationnel. D'autres critères (indice d'impureté de Gini, mesure d'incertitude de Fusinter…) tentent de concilier les deux aspects en étant à la fois sensible aux effectifs et à la distribution de la variable à prédire. Le critère MDL (Minimum Description Length) (Fayyad 1992) est une approche originale qui cherche à optimiser la quantité totale d'information contenue dans le modèle et les exceptions au modèle.

Nous présentons une nouvelle méthode de discrétisation appelée Khiops. Il s'agit d'une méthode ascendante basée sur l'optimisation globale du Khi2. Les méthodes existantes les plus proches sont les méthodes descendantes et ascendantes utilisant le critère du Khi2, mais de façon locale. La méthode descendante basée sur le Khi2 est ChiSplit. Elle recherche le meilleur point de coupure d'un intervalle, en maximisant le critère du Khi2 appliqué aux deux sous-intervalles de part et d'autre du point de coupure : on coupe un intervalle si les deux sous-intervalles présentent des différences significatives statistiquement. Le critère d'arrêt est une probabilité d'indépendance maximum à respecter (calculée d'après la loi du Khi2). La méthode ascendante basée sur le Khi2 est ChiMerge (Kerber 1991). Elle recherche la meilleure fusion d'intervalles adjacents en minimisant le critère du Khi2 : on fusionne deux intervalles adjacents s'ils sont similaires statistiquement. Le critère d'arrêt est une probabilité d'indépendance minimum à respecter (calculée d'après la loi du Khi2).

La méthode Khiops commence la discrétisation à partir des intervalles élémentaires réduits à un individu. Elle évalue toutes les fusions d'intervalles adjacents et choisit celle qui maximise le critère du Khi2 appliqué à la distribution de l'ensemble des intervalles. Le critère d'arrêt est basé sur la probabilité d'indépendance associée au Khi2. La méthode s'arrête automatiquement dès que la probabilité d'indépendance ne décroît plus. La méthode Khiops optimise un critère d'évaluation global de la partition du domaine en intervalles, et non un critère local appliqué à deux intervalles adjacents comme dans ChiSplit ou ChiMerge. Son absence complète de paramétrage la rend très souple à utiliser et permet d'aboutir à des partitions de grande qualité sans intervention utilisateur. En dépit de cette approche globale, l'algorithme associé à la méthode Khiops est en Nlog(N) ou N est le nombre d'instances à discrétiser. Cette complexité algorithmique est la même que pour l'algorithme ChiMerge optimisé.

Le partitionnement réalisé lors d'une discrétisation constitue un modèle prédictif élémentaire basé sur un seul attribut descriptif, en prédisant dans chaque intervalle la modalité cible majoritaire. Une méthode de discrétisation peut ainsi être considérée comme un algorithme inductif, soumis intrinsèquement au risque de sur-apprentissage. Bien que connu, ce problème de sur-apprentissage n'a pas été analysé ni pris en compte de façon approfondie pour le problème de la discrétisation supervisée. Dans le cas de l'algorithme Khiops, un effectif minimum par intervalle permet de réduire le risque de sur-apprentissage. La valeur choisie pour cet effectif minimum est la racine carrée de la taille de l'ensemble d'apprentissage. Ce choix empirique permet d'assurer à la fois une amélioration de la fiabilité statistique (effectif des intervalles) et de la valeur informationnelle (nombre d'intervalles potentiels) quand la taille de l'ensemble d'apprentissage augmente. Nous proposons ici une amélioration significative de l'algorithme, qui permet un contrôle réel (et non plus empirique) de la qualité des discrétisations. Le principe utilisé repose sur l'analyse du comportement de l'algorithme en présence d'un attribut numérique indépendant de l'attribut à prédire. Nous avons étudié la statistique de la variation du Khi2 lors de la fusion de deux intervalles, puis modélisé l'algorithme Khiops en termes de fusions d'intervalles, et enfin proposé une estimation de la variation maximale du Khi2 obtenue lors de la discrétisation d'un attribut indépendant de l'attribut à prédire. Cette estimation, vérifiée expérimentalement, permet alors de modifier l'algorithme Khiops en le contraignant à accepter toute fusion d'intervalle entraînant une variation du Khi2 inférieure à la variation maximale théorique calculée. Cette amélioration permet de garantir d'une part que les discrétisations d'un attribut indépendant de l'attribut à prédire aboutissent à un seul intervalle terminal, d'autre part que les discrétisations aboutissant à plusieurs intervalles correspondent à des attributs ayant un intérêt prédictif réel.

Afin d'évaluer la méthode Khiops et de la comparer à plusieurs autre méthode de discrétisation, nous avons procédé à des expérimentations intensives sur des bases de test provenant de l'UCI Irvine (Blake 1998). Nous avons pris en compte plusieurs critères d'évaluation, qui sont principalement la performance prédictive, la robustesse (dégradation de la performance entre apprentissage et test), la taille des discrétisations et le comportement vis à vis des attributs bruités. Une analyse multi-critères des résultats permet de comparer en détail les différentes méthodes et la sensibilité à leur paramétrage. Cette évaluation en profondeur montre que la méthode Khiops est notablement performante sur l'ensemble des critères considérés.

Le reste du document est organisé de la façon suivante.

La partie 2 présente l'algorithme Khiops initial et ses propriétés fondamentales. La partie 3 compare la méthode Khiops avec les méthodes apparentées ChiMerge et ChiSplit d'un point de vue théorique. La partie 4 présente les améliorations apportées à la méthode, notamment en se basant sur une analyse de la statistique du comportement de l'algorithme. La partie 5 procède à des expérimentations comparatives permettant une évaluation multi-critères des méthodes de discrétisation. L'annexe étudie les problèmes de sensibilité numérique liés à l'approximation de la loi du Khi2.

2 La méthode de discrétisation Khiops initiale

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 Chi2

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 $e_{ij} = n_{i.} * n_{.j} / N$. e_{ij} représente 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.

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

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 Algorithme Khiops

Le test du Khi2 est à la fois sensible aux effectifs et aux proportions des modalités cibles. Il s'agit donc d'un critère intéressant a priori pour les méthodes de discrétisation. 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 va chercher 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 dans chaque cellule du tableau de Khi2. Cette contrainte devra être prise en compte dans l'optimisation.

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

Algorithme Khiops

- Initialisation
  - Tri des valeurs de la loi source
  - Création d'un intervalle élémentaire par valeur de la loi source
  - Calcul de la probabilité d'indépendance entre la loi discrétisée et la loi cible
- Optimisation de la discrétisation
  - Répéter
    - Evaluer toutes les fusions possibles d'intervalles adjacents
      - ✓ Calcul du Khi2 associé à la nouvelle loi discrétisée 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).

2.3 Effectif minimum par intervalle

La convention la plus courante est d'exiger que les effectifs théoriques soient au moins égaux à 5 pour chaque case du tableau de contingence. Cette convention doit être respectée pour des raisons de fiabilité de la loi du Khi2. Cet effectif théorique minimum par case est équivalent à un effectif minimum par ligne du tableau du Khi2, et donc à un effectif minimum par intervalle de la discrétisation.

Dans le cadre de la discrétisation, on procède à des regroupements de valeurs adhoc en espérant approximer les proportions des modalités cibles à partir des régularités observées dans l'échantillon. Ces régularités proviennent en fait non seulement de la loi de distribution, mais également du hasard lié à l'échantillon. Afin de ne pas se baser à tort sur des régularités qui proviendraient uniquement du hasard, c'est à dire de "sur-apprendre" l'échantillon, une solution est d'augmenter la valeur de l'effectif minimum par intervalle, afin de lisser les effets du hasard. On prendra pour valeur de l'effectif minimum par intervalle la racine carrée de la taille de l'échantillon. Cette valeur permet d'une part d'améliorer la fiabilité statistique de l'évaluation de la loi de distribution sur chaque intervalle discrétisé, d'autre part d'augmenter le nombre d'intervalles potentiels et donc la finesse de la discrétisation quand la taille de l'échantillon augmente.

En définitive, on prendra pour effectif minimum par intervalle le maximum du résultat des deux calculs précédents pour assurer à la fois la fiabilité statistique du test du Khi2 et prévenir les problèmes de sur-apprentissage.

2.4 Exemple

On va illustrer le déroulement de l'algorithme sur la base Iris provenant des bases d'apprentissage de l'UCI Irvine (Blake 1998). La base Iris est composée de 150 instances. Les instances, représentant des fleurs de la famille des Iris, sont décrites par 5 attributs :

- sepal length en cm
- sepal width en cm
- petal length en cm
- petal width en cm
- class: Iris setosa, Iris versicolor, Iris virginica

La variable à prédire est la classe. On va discrétiser l'attribut sepal width, qui étant le moins corrélé avec la variable cible est le plus intéressant pour illustrer la méthode. Le tableau de contingence associé aux valeurs de l'attribut sepal width est décrit dans le tableau 2.

Tableau 2 : Tableau de contingence pour l’attribut sepal width de la base Iris. Evaluation des fusions.

Sepal width	Iris versicolor	Iris virginica	Iris setosa	Total
2.0	1	0	0	1
2.2	2	1	0	3
2.3	3	0	1	4
2.4	3	0	0	3
2.5	4	4	0	8
2.6	3	2	0	5
2.7	5	4	0	9
2.8	6	8	0	14
2.9	7	2	1	10
3.0	8	12	6	26
3.1	3	4	5	12
3.2	3	5	5	13
3.3	1	3	2	6
3.4	1	2	9	12
3.5	0	0	6	6
3.6	0	1	2	3
3.7	0	0	3	3
3.8	0	2	4	6
3.9	0	0	2	2
4.0	0	0	1	1
4.1	0	0	1	1
4.2	0	0	1	1
4.4	0	0	1	1
Total	50	50	50	150

Merged Interval	Chi-square merge value
] -∞ ; 2.25]	87.86
]2.10; 2.35]	87.44
]2.25; 2.45]	87.72
]2.35; 2.55]	85.09
]2.45; 2.65]	88.18
]2.55; 2.75]	88.33
]2.65; 2.85]	87.83
]2.75; 2.95]	84.49
]2.85; 3.05]	83.18
]2.95; 3.15]	87.03
]3.05; 3.25]	88.29
]3.15; 3.35]	88.12
]3.25; 3.45]	84.86
]3.35; 3.55]	87.20
]3.45; 3.65]	87.03
]3.55; 3.75]	87.36
]3.65; 3.85]	87.03
]3.75; 3.95]	87.36
]3.85; 4.05]	88.36
]3.95; 4.15]	88.36
]4.05; 4.25]	88.36
]4.15; +∞ [	88.36

Lors de l’initialisation, on constitue les 23 intervalles élémentaires]- ∞; 2,1],]2,1; 2,25] …]4,15; 4,3],]4,3; +∞[. La valeur du Khi2 associée est de 88,36. En prenant la loi du Khi2 à 44 degrés de liberté correspondante (44=(23-1)*(3-1)), on obtient une probabilité d’indépendance de 8,3 10^{-5}. On calcule alors le Khi2 résultant de chaque fusion d’intervalles. Par exemple, la fusion des intervalles]-∞; 2,1],]2,1; 2,25] donne un nouvel intervalle]-∞; 2,25] et le Khi2 résultant de la nouvelle table (avec un intervalle en moins) a une valeur de 87,86. On cherche alors la fusion qui maximise le Khi2. Ici, la valeur max du Khi2 résultant d’une fusion est de 88,36, atteinte par exemple pour la fusion des deux derniers intervalles]4,15; 4,3] et]4,3; +∞[. En prenant la loi du Khi2 à 42 degrés de liberté correspondante (il y a un intervalle en moins), on obtient une probabilité d’indépendance de 3,8 10^{-5}. La probabilité d’indépendance diminuant, la discrétisation est améliorée et on réalise la fusion correspondante. On recommence ces étapes tant qu’il y a amélioration de la discrétisation.

Le tableau 3 illustre la liste des étapes successive de la méthode de discrétisation. Pour chaque intervalle constitué, on a rappelé les effectifs observés correspondants. Au départ, les intervalles sont fusionnés pour arriver à respecter la contrainte des effectifs minimaux par intervalle, tout en optimisant le critère de discrétisation. Une fois la contrainte satisfaite, les fusions d’intervalles se font uniquement pour optimiser le critère de discrétisation. Comme les trois modalités cibles sont équidistribuées, il faut un effectif ligne observé de 15 pour satisfaire la contrainte d’effectif théorique par case de 5. Cette valeur étant supérieure à racine de 150 (contrainte pour éviter le sur-apprentissage), on utilise ici un effectif minimum par intervalle de 15.

Tableau 3 : Fusions successives des intervalles pour arriver à une discrétisation en trois intervalles

Sepal width	Iris versicolor	Iris virginica	Iris setosa	Total					
2.0	1	0	0	1	3-1-0	9-1-1			34-21-2
2.2	2	1	0	3					
2.3	3	0	1	4	6-0-1				
2.4	3	0	0	3					
2.5	4	4	0	8		12-10-0	18-18-0	25-20-1	
2.6	3	2	0	5	8-6-0				
2.7	5	4	0	9					
2.8	6	8	0	14					
2.9	7	2	1	10					
3.0	8	12	6	26			15-24-18		
3.1	3	4	5	12	6-9-10	7-12-12			
3.2	3	5	5	13					
3.3	1	3	2	6					
3.4	1	2	9	12	1-2-15		1-5-24	1-5-30	
3.5	0	0	6	6					
3.6	0	1	2	3	0-1-5	0-3-9			
3.7	0	0	3	3					
3.8	0	2	4	6					
3.9	0	0	2	2			0-0-6		
4.0	0	0	1	1	0-0-2	0-0-4			
4.1	0	0	1	1					
4.2	0	0	1	1	0-0-2				
4.4	0	0	1	1					
Total	50	50	50	150					

Au bout d'une vingtaine d'étapes, on arrive à la loi discrétisée du tableau 4:

Tableau 4: Tableau de contingence pour l'attribut sepal width discrétisé de la base Iris

Sepal width	Iris Versicolor	Iris virginica	Iris setosa	Total
] -∞ ; 2.95[	34	21	2	57
[2.95; 3.35[	15	24	18	57
[3.35; +∞ [	1	5	30	36
Total	50	50	50	150

Merged Interval	Chi-square Merge value
] -∞ ; 3.35]	54.17
]2.95; +∞]	43.97

Le Khi2 associé à la loi discrétisée a une valeur de 70,74, ce qui correspond à une probabilité d'indépendance de 1,66 10^{-14} (loi du Khi2 à 4 degrés de liberté). Deux fusions d'intervalles sont encore possibles. La meilleure d'entre elles est la première fusion, qui correspond à un Khi2 de valeur 54,17. La probabilité d'indépendance associée est 1,73 10^{-12} (loi du Khi2 à 2 degrés de liberté). Cette fusion qui entraîne une croissance de la probabilité d'indépendance est donc refusée.

La variable sepal width a donc été discrétisée en trois intervalles. Dans le premier intervalle, la classe Iris setosa est très rare. Dans le second, il y a équilibre entre les trois classes. Dans le dernier intervalle, la classe Iris setosa est de loin la plus fréquente.

2.5 Complexité algorithmique

On va évaluer la complexité algorithmique de la méthode de discrétisation Khiops par rapport au nombre d'instances N de la base de données de travail. Dans le pire des cas, les instances prennent des valeurs toutes différentes pour la variable à discrétiser.

Si l'on se base sur les étapes de l'algorithme Khiops, on obtient une complexité algorithmique en N^3.

- Initialisation: en Nlog(N)
- Optimisation de la discrétisation
  - Répéter (au plus N étapes)
    - Evaluer toutes les fusions possibles d'intervalles adjacents : N évaluation de Khi2 (en N)
    - Chercher la meilleure fusion : en N
    - Evaluer la condition d'arrêt : en 1

On va montrer que l'on peut optimiser l'algorithme et le ramener à une complexité algorithmique en Nlog(N). Le calcul du Khi2 sur un tableau de contingence complet demande N étapes de calcul de Khi2 ligne.

$$Khi2 = \sum_i Khi2l_i$$

Le calcul du Khi2 correspondant à la fusion de deux lignes i et i' (i'=i+1) peut s'écrire de la façon suivante :

$$Khi2F_{ii'} = \sum_{k<i} Khi2l_k + Khi2l_{ii'} + \sum_{k>i'} Khi2l_k$$

$$Khi2F_{ii'} = \sum_{k} Khi2l_k + Khi2l_{ii'} - Khi2l_i - Khi2l_{i'}$$

$$Khi2F_{ii'} = Khi2 + DeltaKhi2_{ii'}$$

Grâce à l'additivité du critère du Khi2, le Khi2 lié à une fusion d'intervalles peut être évalué en une seule étape si l'on connaît le Khi2 initial. Si l'on mémorise toutes les valeurs de Khi2 ligne et de DeltaKhi2, la recherche de la meilleure fusion se fait en recherchant le meilleur DeltaKhi2. Après une fusion d'intervalles, seuls les intervalles adjacents à l'intervalle fusionné doivent être mis à jour pour préparer l'étape suivante. La partie critique de l'algorithme devient alors la recherche de la meilleure fusion à chaque étape. Cette recherche est en N. Si l'on trie préalablement la liste des fusions possibles, et que l'on maintient cette liste triée au cours de l'optimisation de la discrétisation, la recherche du meilleur élément est en 1, au prix du coût de gestion de la liste triée. Les arbres binaires de recherche équilibrés (AVL Binary Search Tree par exemple) permettent de gérer une telle liste triée en maintenant l'ordre dans la liste lors d'insertions/suppressions à un coût logarithmique.

En se basant sur la mémorisation des Khi2Ligne et des DeltaKhi2, sur le calcul incrémental des Khi2 et sur l'utilisation d'une liste triée de type arbre binaire de recherche équilibré, on arrive alors à une complexité globale de Nlog(N).

Algorithme Khiops optimisé

- Initialisation
  - Tri des valeurs de la loi source : en Nlog(N)
  - Création d'un intervalle élémentaire par valeur de la loi source : en N
  - Calcul des Khi2 ligne et du Khi2 initial : en N
  - Calcul des DeltaKhi2 : en N
  - Tri des fusions par valeur de DeltaKhi2 : en Nlog(N)
  - Calcul de la probabilité d'indépendance entre la loi discrétisée et la loi cible : en 1
- Optimisation de la discrétisation
  - Répéter: N étapes
    - Chercher la meilleure fusion : en 1 en prenant le premier élément de la liste triée
    - 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)
    - Si continuer : effectuer la fusion d'intervalle
      - Calcul du Khi2Ligne pour le nouvel intervalle : en 1
      - Calcul des DeltaKhi2 pour les deux intervalles adjacents au nouvel intervalle
      - Mise à jour de la liste triée des DeltaKhi2 : en log(N)
        - ✓ Suppression du DeltaKhi2 du nouvel intervalle
        - ✓ Suppression des anciens DeltaKhi2 des intervalles adjacents aux deux sous intervalles sources du nouvel intervalle
        - ✓ Ajout des nouveaux DeltaKhi2 des intervalles adjacents au nouvel intervalle

On peut noter que l'occupation mémoire nécessaire pour l'algorithme est également en Nlog(N). On doit en effet mémoriser N Khi2 lignes, N DeltaKhi2, et une structure de liste triée de type arbre binaire de recherche équilibré qui a une occupation mémoire de Nlog(N).

La version optimisée de l'algorithme Khiops a la même complexité que la version optimisée de l'algorithme ChiMerge, ce qui rend la méthode utilisable y compris sur des bases de données très volumineuses.

3 Comparaison théorique avec les méthodes basées sur le Khi2

3.1 Propriétés des fusions d'intervalles pour la méthode Khiops

Soit une distribution des modalités cible p_1, p_2, $\ldots p_J$. Soit une première ligne de Khi2, d'effectif *n*, pour des proportions de modalités cibles a_j. Soit une seconde ligne de Khi2, d'effectif *n'*, pour des proportions de modalités cibles b_j.

On a $\sum_j p_j = 1$, $\sum_j a_j = 1$, $\sum_j b_j = 1$.

Les effectifs observés et théoriques sont $a_j n$ et $p_j n$ pour la première ligne, $b_j n'$ et $p_j n'$ pour le second ligne. Les Khi2 lignes sont

$$Khi2l = n\left(\sum_j a_j^2/p_j - 1\right) \text{ et } Khi2l' = n'\left(\sum_j b_j^2/p_j - 1\right).$$

On envisage la fusion des deux lignes de Khi2. Les effectifs observés et théoriques de la ligne fusionnée sont $a_j n + b_j n'$ et $p_j(n+n')$.

Le Khi2 ligne de la fusion est

$$Khi2l'' = (n+n')\left(\sum_j \frac{\left((a_j n + b_j n')/(n+n')\right)^2}{p_j} - 1\right) \qquad [1]$$

Le regroupement des deux lignes entraîne une modification du Khi2, $DeltaKhi2 = Khi2l'' - Khi2l - Khi2l'$.

$$DeltaKhi2 = \sum_j \frac{(n+n')\left((a_j n + b_j n')/(n+n')\right)^2 - n a_j^2 - n' b_j^2}{p_j} \qquad [2]$$

$$\boxed{DeltaKhi2 = -\frac{nn'}{n+n'}\sum_j \frac{(a_j - b_j)^2}{p_j}} \qquad [3]$$

La fusion de deux lignes de Khi2 ne peut que faire décroître la valeur du Khi2. La loi du Khi2 a cependant moins de degrés de liberté. Si le Khi2 décroît suffisamment faiblement (voire ne décroît pas), la probabilité d'indépendance correspondante diminue. Sinon, cette probabilité augmente.

3.2 Comparaison avec ChiMerge

Pour la méthode ChiMerge, on considère le tableau de contingence local aux deux lignes. Dans ce contexte local, la distribution des modalités cibles q_1, q_2, $\ldots q_J$ a pour valeurs $q_j = (a_j n + b_j n')/(n+n')$. Pour évaluer l'intérêt de la fusion des deux lignes, on calcule le Khi2 de cette table locale du Khi2.

$$SommeKhi2l = n\left(\sum_j \frac{a_j^2}{q_j} - 1\right) + n'\left(\sum_j \frac{b_j^2}{q_j} - 1\right) \qquad [4]$$

$$\boxed{SommeKhi2l = \frac{nn'}{n+n'}\sum_j \frac{(a_j - b_j)^2}{q_j}} \qquad [5]$$

Le calcul du critère d'arrêt pour les méthodes Khiops et ChiMerge conduit donc à une expression mathématique similaire. Par contre, l'interprétation du critère est radicalement différente. La distribution des modalités cibles est globale à toute la table pour Khiops (proportions p_i), alors qu'elle est locale aux deux lignes adjacentes de la table pour ChiMerge (proportions q_i).

Pour Khiops, on s'arrête si:

Proba(Khi2+DeltaKhi2, (n-2)(J-1)) < Proba(Khi2, (n-1)*(J-1))*

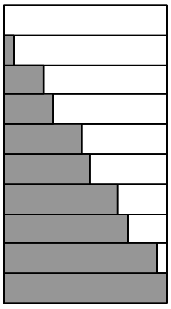
Pour ChiMerge (paramétré par une valeur ProbaSeuil), on s'arrête si:

Proba(SommeKhi2l, J-1) > ProbaSeuil

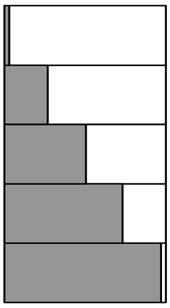
Cela illustre une différence essentielle entre les méthodes. ChiMerge fonctionne de façon locale, alors que Khiops tient compte des proportions de modalités cibles globales, du nombre d'intervalles global et de la valeur globale du Khi2.

Le tableau 5 illustre la difficulté de choisir un seuil de Khi2 pour ChiMerge.

Tableau 5 : Choix de la meilleure fusion d'intervalle pour Khiops et ChiMerge.



Initial table		Khiops	ChiMerge		Final table	
0	100	ΔChi2	LChi2	Prob		
6	94	-0.72	6.19	0.013	6	194
24	76	-6.48	12.71	0.000		
30	70	-0.72	0.91	0.339	54	146
47	53	-5.78	6.10	0.013		
53	47	-0.72	0.72	0.396	100	100
70	30	-5.78	6.10	0.013		
76	24	-0.72	0.91	0.339	146	54
94	6	-6.48	12.71	0.000		
100	0	-0.72	6.19	0.013	194	6



Le tableau de contingence initiale, sur la gauche, représente un échantillon de 1000 instances avec deux modalités cibles équidistribuées. Dans le tableau initial, on a une série de paires d'intervalles ayant des effectifs voisins, et il paraît naturel lors d'une discrétisation de fusionner chaque paire d'intervalles, ce qui correspond au résultat obtenu dans le tableau de contingence final, sur la droite. Les évaluations des fusions d'intervalles, DeltaKhi2 pour Khiops et LocalKhi2 pour ChiMerge sont présentées dans les colonnes centrales.

On a ici un Khi2 total pour le tableau global de 449,2 égal à environ 50 fois le nombre de degrés de liberté. En se référant à l'étude du calcul numérique des DeltaKhi2 en annexe, les fusions de DeltaKhi2 supérieur à -5 sont acceptées, les autres sont refusées. Pour l'algorithme Khiops, les cinq fusions « évidentes » sont acceptées et considérées comme équivalentes. Pour ChiMerge, les fusions centrales (autour de p=0,5) sont largement préférées aux fusions extrêmes (p = 0,03 ou 0,97). La fusion entre les lignes 30-70 et 47-53 est même préférée à la fusion entre les lignes 0-100 et 6-94. Dans ce cadre, il est difficile de choisir le bon seuil pour l'algorithme ChiMerge.

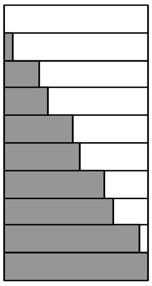
En conclusion, la méthode ChiMerge comporte plusieurs faiblesses intrinsèques qui sont résolues par la méthode Khiops. Les caractéristiques purement locales de ChiMerge entraînent des difficultés pour trouver un paramétrage du seuil de Khi2 optimal. Tout seuil fixé par l'utilisateur ne sera pertinent qu'à certaines étapes de l'algorithme (problèmes d'échelles liées à la taille de l'échantillon initial et au nombre d'intervalles) et avantagera à tort les fusions d'intervalles dont les proportions locales sont proches de l'équipartition. Le critère global utilisé dans Khiops résout ces problèmes en calculant un critère d'arrêt auto-adaptatif en fonction de la taille de l'échantillon et des spécificités locales des intervalles évalués équitablement parmi l'ensemble de toutes les fusions possibles.

3.3 Comparaison avec ChiSplit

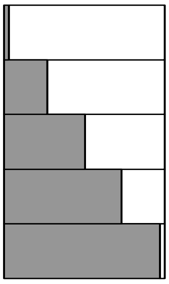
Khiops est un algorithme ascendant et ChiSplit est un algorithme descendant, ce qui rend la comparaison entre les deux méthodes plus difficile que pour ChiMerge. Le critère d'arrêt de ChiSplit est très délicat à ajuster car il dépend de facteurs d'échelle (nombre de lignes du tableau), de l'importance des singularités de l'attribut à discrétiser, et de la position de ces singularités dans la table du Khi2.

On va reprendre le premier exemple utilisé pour ChiMerge pour illustrer l'ensemble de ces problèmes dans le tableau 6.

Tableau 6 : Choix de la meilleure fusion d'intervalle pour Khiops et ChiSplit



Initial table		Khiops	ChiSplit		Final table	
0	100	ΔChi2	Chi2S	Prob		
6	94	-0.72	111.11	5.59E-26	6	194
24	76	-6.48	220.90	5.76E-50		
30	70	-0.72	274.29	1.32E-61	54	146
47	53	-5.78	326.67	5.11E-73		
53	47	-0.72	327.18	3.95E-73	100	100
70	30	-5.78	326.67	5.11E-73		
76	24	-0.72	274.29	1.32E-61	146	54
94	6	-6.48	220.90	5.76E-50		
100	0	-0.72	111.11	5.59E-26	194	6



On est ici dans des ordres de grandeur de 10^{-25} à 10^{-75} pour le seuil de Khi2 à utiliser. Pour des échantillons de taille supérieure (de l'ordre de 10000 individus ou plus), on se retrouverait aux limites de la précision numérique des machines (de l'ordre de 10^{-300}), ce qui rendrait impossible le choix d'un seuil. Par ailleurs, la coupure optimale trouvée par ChiSplit est de découper au milieu du tableau du Khi2. En effet, cette coupure donne deux lignes d'effectifs 107-393 et 393-107, qui constitue une excellente coupure de l'ensemble en deux intervalles. Mais de ce fait, la coupure a séparé irrémédiablement les lignes 47-53 et 53-47 qui seraient intuitivement à fusionner. L'approche de l'algorithme ChiSplit qui combine recherche des structures globales et algorithme glouton présente donc des faiblesses qui peuvent gêner l'identification des régularités locales de la variable à discrétiser.

4 Améliorations de la méthode

4.1 Limites de la méthode

Il faut dissocier la méthode de l'algorithme et de son implémentation. Le principe de la méthode est de rechercher parmi tous les regroupements en intervalles possibles celui qui minimise la probabilité d'indépendance entre la loi discrétisée et la loi cible. Cette probabilité est mesurée par la loi du Khi2 appliquée au tableau de contingence entre loi discrétisée et loi cible. Pour améliorer la fiabilité statistique de l'algorithme, un effectif minimum dépendant de la taille de l'échantillon est ajouté pour contraindre la recherche de la meilleure partition en intervalles. A ce niveau de principe, la méthode Khiops paraît robuste.

L'algorithme nécessite une bonne approximation de la loi du Khi2 pour des valeurs très importantes de nombre de degrés de liberté et de Khi2. L'évaluation exacte de la loi du Khi2 serait l'idéal, mais elle n'est pas disponible dans la pratique. De plus, on arrive aux limites de la précision numérique des ordinateurs pour des probabilités d'indépendance proches de zéro.

Le calcul de l'effectif minimal par intervalle résulte d'un choix heuristique, et n'a pas de fondement solide. Théoriquement, ce calcul devrait s'appuyer sur une estimation statistique prenant en compte précisément la distribution des modalités cibles et contrôlant la probabilité de sur-apprentissage.

L'algorithme de recherche est un algorithme glouton qui prend en compte la contrainte d'effectif minimum de la façon la plus souple possible. Cette heuristique garantit un temps d'exécution super-linéaire, ce qui est indispensable dès que l'on s'attaque à des problèmes de Data Mining tirés du monde réel. Par contre, il est clair que l'algorithme ne conduit pas forcément à la solution optimale et que l'on peut même construire des exemples le mettant en défaut, notamment en ce qui concerne la prise en compte des contraintes d'effectif minimum. Il est néanmoins inenvisageable de rechercher la solution optimale du problème de la discrétisation optimale.

Les limites de la méthode proviennent d'avantage de son implémentation que de son principe. Le problème critique de l'évaluation de la loi du Khi2 a été résolu pour la méthode Khiops initiale. Nous présentons en annexe des méthodes numériques permettant d'approximer le logarithme de la probabilité associée au Khi2 et de calculer de façon très précise les variations du Khi2 contrôlant le critère d'arrêt de l'algorithme Khiops, et ce pour de très larges domaines de valeurs.

La limite la plus importante est l'utilisation de l'effectif minimum pour lutter contre le sur-apprentissage. L'effectif utilisé, égal à la racine carrée de la taille de l'échantillon, est à la fois trop important pour permettre d'isoler des petits intervalles très riches en information, et trop faible pour offrir une véritable garantie contre le sur-apprentissage. Nous allons illustrer ce dernier point en rapportant les résultats d'une expérimentation, qui consiste à discrétiser un attribut indépendant de l'attribut cible. Cette expérimentation montre que les discrétisations aboutissent à plusieurs intervalles et à un indicateur de qualité de la discrétisation *ProbLevel* non nul, ce qui laisse supposer à tort que l'attribut discrétisé contient de l'information sur l'attribut cible. Cela traduit un sur-apprentissage, d'autant plus important que la taille de l'échantillon d'apprentissage est élevée. Le tableau 7 présente les résultats (en moyenne) de 100 discrétisations d'attributs indépendants de l'attribut cible, pour différentes tailles d'échantillon.

Tableau 7 : Résultats de discrétisation d'un attribut descriptif indépendant de l'attribut cible

Taille	Nombre d'intervalles	ProbLevel
100	2,4	1,46
1000	4,1	2,33
10000	8,2	4,59
100000	17,6	9,50
1000000	37,3	20,36

La méthode Khiops ne permet donc pas de définir un niveau « plancher » en nombre d'intervalles ou en valeur de ProbLevel correspondant aux attributs indépendants de l'attribut cible. Le choix empirique de l'effectif minimum n'est donc pas satisfaisant en présence d'attributs sans intérêt prédictif. De plus, il ne tient pas compte du nombre et de la distribution des modalités cibles.

4.2 Amélioration par analyse de la statistique de l'algorithme

4.2.1 Présentation

L'algorithme Khiops envisage toutes les fusions possibles d'intervalles, choisit la meilleure fusion, et si le critère d'arrêt n'est pas atteint, effectue cette fusion et réitère ces opérations. Dans le cas où l'attribut descriptif et l'attribut cible sont indépendants, le résultat souhaité serait que Khiops aboutissent à un seul groupe terminal, c'est à dire à la conclusion qu'il n'y a aucune valeur informationnelle dans l'attribut descriptif.

Pour deux attributs indépendants, la valeur du Khi2 suit une loi de probabilité dont l'espérance et la variance sont connues. On va de même étudier la loi du DeltaKhi2 (variation de la valeur du Khi2 lors de la fusion de deux intervalles) dans le cas de deux attributs indépendants. Lors du déroulement de l'algorithme Khiops, un grand nombre de fusions sont envisagées, et Khiops à chaque étape choisit la meilleure de toutes ces fusions en optimisant le critère du Khi2, ou ce qui est équivalent en optimisant le critère du DeltaKhi2 (le Khi2 de départ est fixé). Khiops arrête les fusions quand la valeur du meilleur DeltaKhi2 devient trop importante. Pour deux attributs indépendants, il faut cependant continuer les fusions jusqu'à ce qu'il ne reste qu'un groupe final égal à l'échantillon initial. Il faut alors que le plus grand DeltaKhi2 rencontré au cours de l'algorithme soit accepté. On va essayer d'estimer la valeur de ce MaxDeltaKhi2 au cours de l'algorithme, et imposer que les fusions soient continuées tant que ce seuil n'est pas atteint.

4.2.2 Loi du DeltaKhi2

Pour une loi du Khi2 à k degrés de liberté et un échantillon de taille N, on a l'espérance et la variance suivante :

$$E(Khi2)=k$$

$$Var(Khi2) = 2k + \frac{1}{N}\left(\sum_{i=1}^{k}\frac{1}{q_i} - k^2 - 4k - 1\right)$$

On rappelle ci dessous la formule du DeltaKhi2 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
	…	…	…	…	…
	…	…	…	…	
	$p_1 n$	$p_2 n$	…	$p_J n$	n
	$p'_1 n$	$p'_2 n$	…	$p'_J n'$	n'
	…	…	…	…	
	…	…	…	…	
Total	$P_1 N$	$P_2 N$	…	$P_J N$	N

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

Cette variation est toujours négative, et n'est nulle que si les intervalles 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.

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

Le calcul de la fonction de répartition de DeltaKhi2 est basé sur des lois binomiales discrètes, ce qui le rend difficile à évaluer pour des valeurs importantes de n. On va utiliser le théorème central limite pour approximer la loi du DeltaKhi2 dans le cas où n=n'.

Proposition : Pour un attribut descriptif indépendant d'un attribut cible à J modalités, le DeltaKhi2 de la fusion de deux intervalles de même effectif suit asympotiquement une loi du Khi2 à J-1 degrés de liberté.

Preuve :

On va donner la preuve de ce résultat dans le cas de deux modalités cibles, et admettre sa validité dans les autres cas.

Dans le cas de deux modalités cibles, posons $P=P_1=1-P_2$, $p=p_1=1-p_2$, $p=p'_1=1-p'_2$.

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

$DeltaKhi2 = \frac{n}{2P(1-P)}\left(\frac{k}{n} - \frac{k'}{n}\right)^2$ où k et k' suivent des lois binomiales de paramètre p et p'.

Soit X=Bernouilli(p)-Bernouilli(p').

Soit $S_n = k - k'$. S_n est la somme de n variables de type X, identiquement distribuées et indépendantes.

E(X)=p-p' V(X)=p(1-p) + p'(1-p')

Alors $\frac{\sqrt{n}}{\sqrt{V(X)}}\left(\frac{S_n}{n} - E(X)\right)$ converge en loi vers la loi de Gauss centrée réduite.

Dans l'hypothèse d'indépendance des attributs, on a p=p'=P.

Dans ce cas, $DeltaKhi2 = \left(\frac{\sqrt{n}}{\sqrt{V(X)}}\left(\frac{S_n}{n} - E(X)\right)\right)^2$

DeltaKhi2 se comporte comme le carré d'une loi normale, c'est à dire comme la loi du Khi2 à 1 degré de liberté.

Corollaire : Pour un attribut descriptif indépendant d'un attribut cible à J modalités, le DeltaKhi2 de la fusion de deux intervalles de même effectif a les caractéristiques suivantes :

$$p(DeltaKhi2_J \geq DK) \sim p(Khi2_{J-1} \geq DK)$$
$$E(DeltaKhi2_J) \sim J-1$$
$$V(DeltaKhi2_J) \sim 2(J-1)$$

On peut remarquer que la loi du DeltaKhi2 dépend du nombre de modalités cibles, mais pas de leur distribution.

4.2.3 Statistique des fusions de l'algorithme Khiops

Lors d'une discrétisation « totale » jusqu'à un seul intervalle final, le nombre de fusions effectuées est égal à la taille N de l'échantillon (exactement N-1 fusions). Une modélisation simple de l'algorithme Khiops est que toutes ces fusions sont équiprobables et qu'elles suivent la loi du DeltaKhi2. Cette modélisation est extrêmement grossière pour les raisons suivantes :

- ✓ Les fusions envisagées ne sont pas indépendantes
- ✓ L'approximation de la loi du DeltaKhi2 n'a été calculée que pour des intervalles d'effectifs égaux, asymptotiquement (pour des effectifs suffisamment grands).
- ✓ L'algorithme impose une contrainte d'effectifs minimum théorique de 5 par case du tableau de contingence pour les fusions au départ
- ✓ A chaque étape, la fusion effectuée est la meilleure parmi toutes les fusions d'intervalles restant
- ✓ …

On va dans un premier temps évaluer expérimentalement le comportement réel de l'algorithme pour évaluer cette modélisation statistique de Khiops. L'expérimentation consiste à lancer l'algorithme Khiops sur un échantillon comportant un attribut descriptif continu indépendant de l'attribut à prédire, comportant deux modalités cibles équidistribuées. On a au cours du déroulement de l'algorithme effectué toutes les fusions jusqu'à un intervalle terminal unique (sans critère d'arrêt) et collecté les valeurs de DeltaKhi2 de ces fusions afin d'en tracer la fonction de répartition. On effectue cette expérimentation sur des échantillons de taille 100, 1000 et 10000, puis on compare les fonctions de répartition obtenues à la fonction de répartition théorique du DeltaKhi2 de deux intervalles de même effectifs (loi du Khi2 à un degré de liberté). Les résultats sont présentés sur la figure 1.

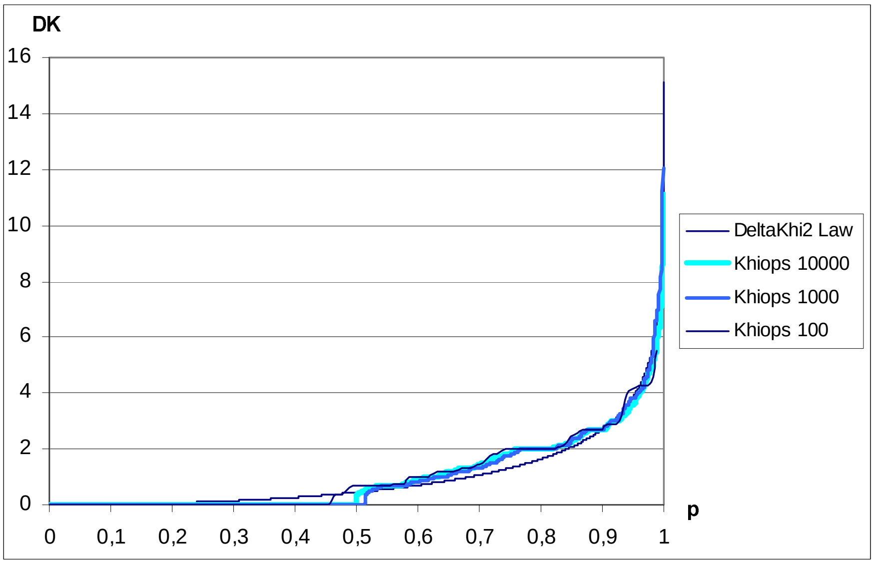


Figure 1 : Loi des DeltaKhi2 des fusions d'intervalles lors d'une exécution de l'algorithme Khiops

L'expérimentation montre que la loi des DeltaKhi2 des fusions effectuées lors d'une exécution de l'algorithme Khiops ne dépend pas de la taille de l'échantillon, et est bien modélisée par la loi théorique du DeltaKhi2 de deux intervalles de même effectif. En début de courbe, les DeltaKhi2 se trouvent sous la courbe théorique. Cela correspond au grand nombre de fusions initiales d'intervalles élémentaires identiques (fusion d'intervalles 0-1 et 0-1, ou d'intervalles 1-0 et 1-0), dont le DeltaKhi2 vaut 0. Après une progression jusqu'à la valeur 2, on a un palier assez important pour DK=2, qui correspond au nombre important de fusions entre intervalles 0-1 et 1-0, imposées pour des contraintes d'effectif minimum par intervalle. Ce début de courbe suit assez mal la courbe théorique : on a ici les effets conjugués de la contrainte d'effectif minimum de l'algorithme, et des fusions opérant sur des intervalles de très petite taille, dans un domaine où la modélisation du DeltaKhi2 par la loi du Khi2 n'est qu'approximative. Après ce palier (à partir de p~0,85), les courbes obtenues lors de l'expérimentation semblent suivre fidèlement la courbe théorique.

4.2.4 Statistique du MaxDeltaKhi2 de l'algorithme Khiops

On cherche un seuil MaxDeltaKhi2 pour l'algorithme Khiops, tel que pour deux attributs indépendants, l'algorithme converge vers un seul groupe terminal avec une probabilité supérieure à p (p = 0,95 par exemple). Il faut donc que toutes les fusions envisagées soient acceptées, c'est à dire que tous les DeltaKhi2 des fusions envisagées soient inférieurs au seuil MaxDeltaKhi2. En se basant sur la modélisation précédente où toutes les fusions sont indépendantes, la probabilité que toutes les fusions envisagées soit acceptées est égale à la probabilité qu'une fusion soit acceptée à la puissance N.

On cherche donc MaxDeltaKhi2 tel que :

$$P(DeltaKhi2_I \leq MaxDeltaKhi2)^N \geq p$$

En passant par la loi du Khi2 équivalente, on a :

$$P(Khi2_{J-1} \leq MaxDeltaKhi2) \geq p^{1/N}$$

$$MaxDeltaKhi2 = InvKhi2_{J-1}(prob \geq p^{1/N})$$

Pour valider cette modélisation de la loi du MaxDeltaKhi2 de l'algorithme Khiops, on s'intéresse cette fois non pas à la distribution des valeurs du DeltaKhi2 au cours d'une exécution de l'algorithme Khiops, mais au max de ces valeurs. Pour cela, on utilise des échantillons de deux attributs indépendants comme précédemment, et on collecte pour un grand nombre d'échantillons à discrétiser la valeur max des DeltaKhi2 des fusions d'intervalles. Cette expérimentation est réalisée 1000 fois pour des échantillons de taille 100, 1000, 10000 et 100000, afin de tracer en figure 2 les fonctions de répartition « empiriques » de MaxDeltaKhi2 pour chacune de ces tailles d'intervalles. On compare ces fonctions de répartition empiriques avec les fonctions de répartition théoriques obtenue avec la formule $MaxDeltaKhi2 = InvKhi2_{J-1}(prob \geq p^{1/N})$.

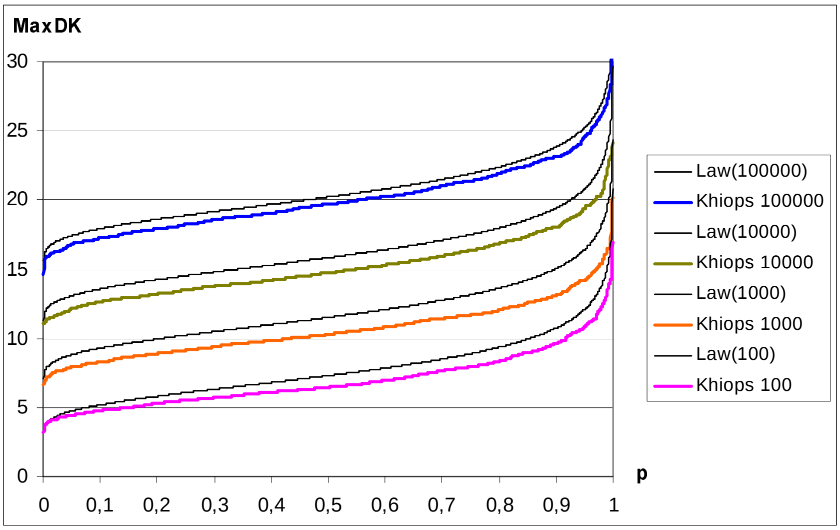


Figure 2 : Lois empiriques et théoriques du MaxDeltaKhi2 de l'algorithme Khiops

On remarque que les lois empiriques et théoriques ont une forme très similaire, quelle que soit la taille de l'échantillon. Les valeurs théoriques constituent une borne supérieure des valeurs empiriques. Cette borne constitue une estimation assez fidèle des valeurs empiriques. Attention, cette borne est le résultat d'une modélisation « grossière » de l'algorithme, et bien que reposant sur des bases raisonnables, son comportement de borne sup n'est vérifié ici qu'expérimentalement.

On a tracé sur la figure 3 la valeur théorique du MaxDeltaKhi2 en fonction de la taille de l'échantillon pour différentes valeurs de p. Ainsi, pour un échantillon de 1000 individus par exemple, alors que le DeltaKhi2 des fusions d'intervalles a une valeur moyenne de 1, la valeur maximale du DeltaKhi2 ne reste inférieure à MDK=8 que dans 1% des cas, ce qui signifie qu'elle dépasse 8 dans 99% des cas. Elle dépasse 15 dans 10% des cas (p=0,9) et 19 dans 1% des cas (p=0,99). Pour un échantillon de taille 1000000, le MaxDeltaKhi2 dépasse 21 dans 99% des cas et 33 dans 1% des cas.

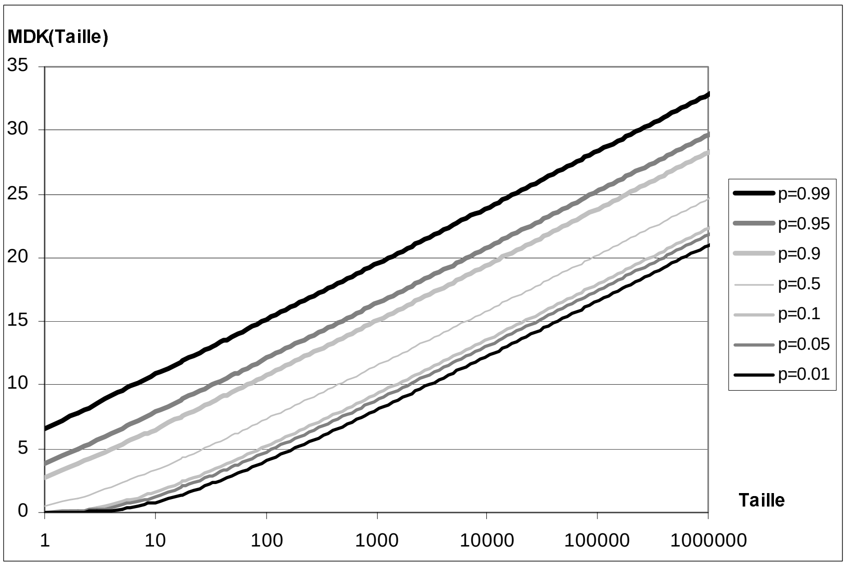


Figure 3 : MaxDeltaKhi2 atteint en fonction de la taille de l'échantillon pour différentes probabilités.

4.2.5 Algorithme Khiops robuste

4.2.5.1 Impact sur l'algorithme Khiops

On rappelle que l'algorithme Khiops effectue des fusions d'intervalles tant qu'il y a diminution de la probabilité d'indépendance entre attribut discrétisé et attribut à prédire, et tant qu'un effectif minimum par intervalle n'est pas atteint. On va garder ici la contrainte d'effectif minimum de 5 par case du tableau du Khi2, qui est justifiée par une évaluation correcte de la statistique du Khi2. Par contre, on ne va plus utiliser la contrainte « empirique » d'effectif minimum en racine carrée de la taille de l'échantillon. Cette contrainte visant à contrer les risques de sur-apprentissage sera remplacée en se basant sur l'étude précédente de la statistique du MaxDeltaKhi2.

Pour deux attributs indépendants, le résultat souhaité est qu'à l'issue de la discrétisation, il ne reste qu'un intervalle, signifiant ainsi que l'attribut prédictif (pris isolément) ne contient pas d'information sur l'attribut à prédire. Dans le cas de deux attributs indépendants, on peut pour une probabilité p donnée déterminer une valeur théorique MaxDeltaKhi2(p) qui ne sera pas dépassée avec une probabilité p. Les expérimentations ayant montré que la loi théorique bornait la loi empirique, cette valeur MaxDeltaKhi2(p) ne sera en pratique pas dépassée avec une probabilité supérieure à p. On impose alors que lors de l'algorithme Khiops, les fusions soient systématiquement acceptées tant que la valeur du DeltaKhi2 est inférieure à MaxDeltaKhi2(p). On assure ainsi le comportement désiré avec une probabilité p. Dans le cas de deux attributs quelconques (non nécessairement indépendants), cette fiabilisation de l'algorithme nous permet d'affirmer que si l'algorithme produit une discrétisation contenant de l'information (au moins deux intervalles), il y a une probabilité supérieure à p pour que l'attribut descriptif soit réellement porteur d'information sur l'attribut à prédire. En pratique, on prendra p=0,95, ce qui d'après les expérimentations précédentes assure une fiabilité de discrétisation dans plus 95% des cas.

Algorithme Khiops robuste :

- Initialisation
  - Tri des valeurs de l'attribut source
  - Création d'un intervalle élémentaire par valeur de l'attribut source
  - Calcul de la probabilité d'indépendance entre attribut discrétisé et attribut cible
  - Calcul de la valeur MaxDeltaKhi2
    - ✓ $MaxDeltaKhi2 = InvKhi2_{J-1}\left(prob \geq 0{,}95^{1/N}\right)$ (N= taille d'échantillon, J = nombre de modalités cibles)
- Optimisation de la discrétisation : répéter
  - Evaluer toutes les fusions possibles d'intervalles adjacents
    - ✓ Calcul du Khi2 associé à la nouvelle loi discrétisée résultant de la fusion
  - Chercher la meilleure fusion
    - ✓ Fusion améliorant le respect des contraintes d'effectif minimum théorique par case en priorité
    - ✓ Fusion maximisant la valeur du Khi2
  - Evaluer la condition d'arrêt
    - ✓ Si au moins une contrainte d'effectif n'est pas respectée, continuer.
    - ✓ Sinon, 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

4.2.5.2 Impact sur l'indicateur de qualité d'une discrétisation

On rappelle que l'indicateur ProbLevel repose sur la probabilité d'indépendance de l'attribut discrétisé, en se basant sur la valeur de Khi2.

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

Etant donné que le Khi2 final est soumis à une forte variance, on propose de redéfinir l'indicateur ProbLevel en retranchant le seuil MaxDeltaKhi2 au Khi2 final.

$ProbLevel = -\log_{10}\left(P\left(Khi2_{final} - MaxDeltaKhi2\right)\right).$

4.2.5.3 Discussion

Le critère d'arrêt de la méthode Khiops robuste suit deux logiques différentes. La première logique consiste à optimiser la discrétisation en minimisant la probabilité d'indépendance entre attribut discrétisé et attribut cible, et conduit à refuser une fusion de deux intervalles significativement différents. La seconde logique consiste à contrôler le sur-apprentissage en minimisant la probabilité que la différence observée de deux intervalles soit due au hasard, ce qui conduit à accepter la fusion de deux intervalles significativement différents. L'affrontement de ces deux logiques est paradoxal, mais s'explique en fait selon le point de vue adoptée. Dans la logique d'optimisation, on prend une décision ponctuelle de fusion, et le critère de décision est justifié localement. Dans la logique de contrôle du sur-apprentissage, on envisage le processus de discrétisation dans sa globalité comme un ensemble important de décisions ponctuelles du premier type. Chacune de ces décisions ponctuelles est justifiée, mais globalement, plus le nombre de ces décisions augmente, plus il devient probable qu'au moins une d'entre elles soit due au hasard et entraîne donc un sur-apprentissage.

On va maintenant évaluer comment les deux critères d'arrêt antagoniste se traduisent mathématiquement. Pour chaque meilleure fusion évaluée, on a calculé la valeur DeltaKhi2 de la variation du Khi2 suite à la fusion. Le critère d'optimisation de la discrétisation demande d'arrêter les fusions si :

$Prob(Khi2-DeltaKhi2, (I-2)*(J-1)) < Prob(Khi2, (I-1)*(J-1))$

Le critère de sur-apprentissage demande de continuer les fusions si :

$DeltaKhi2 < InvKhi2_{I-1}(Prob>=0.95^{1/N})$

En annexe, on a introduit la fonction $DK(x, n, k)=dx \Leftrightarrow Prob(x-dx, n-k) = Prob(x, n)$. Le critère d'optimisation de la discrétisation se réécrit alors :

$DeltaKhi2 > DK(Khi2, I-1, J-1)$

Le critère d'arrêt global est donc :

$DeltaKhi2 > max(DK(Khi2, I-1, J-1), InvKhi2_{I-1}(Prob>=0.95^{1/N}))$

Pour Khi2 grand devant I, on a $DK(Khi2, I-1, J-1) \sim (J-1)ln(1+Khi2/I)$, et pour Khi2~I, $DK(Khi2, I-1, J-1) \sim (J-1)$.

La valeur du Khi2 est bornée par N, donc $DK(Khi2, I-1, J-1) <= (J-1)ln(N)$.

D'après la figure 3, il apparaît clairement que pour J=2, la valeur du MaxDeltaKhi2 est toujours supérieure à $(J-1)ln(N)$, et donc que seul le critère lié au sur-apprentissage est actif. Pour un nombre de modalités cible plus important, la valeur du MaxDeltaKhi2 augmente faiblement, alors que la valeur de $DK(Khi2, I-1, J-1)$ augmente de façon quasiment additive, et finit donc par dépasser le critère dû au sur-apprentissage. Selon le type de problème, l'un ou l'autre des deux critères peut donc devenir prépondérant. Il est à noter que pour le cas très classique de deux modalités cible, la variante robuste de Khiops change en fait radicalement le comportement de l'algorithme. La version basique de l'algorithme continue les fusions tant que la probabilité d'indépendance entre attribut discrétisé et attribut cible diminue, alors que la version robuste continue les fusions d'intervalles tant qu'il est probable que les dissimilarités observées entre intervalles soient dues au hasard.

4.3 Autres améliorations

4.3.1 Ajustement de l'effectif minimum théorique

La valeur du Khi2 suit asymptotiquement une loi du Khi2, et on admet que la fiabilité du test du Khi2 est assurée pour des effectifs minimums théoriques de 5 par cellule du tableau de contingence.

Soient P_1, P_2, .. P_J les probabilités observées des modalités cibles du tableau de contingence et P_{min} la plus petite de ces probabilités. La contrainte d'effectif minimum théorique de 5 par case du tableau de contingence se traduit en une contrainte sur les effectifs par lignes (donc par intervalle) : $N_{line} >= 5/P_{min}$. Si elle assure la fiabilité du test du Khi2, cette contrainte devient très forte dans le cas des probabilités faibles par colonnes, en imposant un effectif minimum par ligne qui peut devenir non négligeable par rapport au nombre total d'instances.

Essayons maintenant d'évaluer la probabilité $p_{k,N}$ d'apparition d'une séquence « pure » de k modalités identiques de probabilité P_{min} dans un échantillon de N instances. Pour un échantillon de k instances, $p_{k,k} = P_{min}{}^{k}$. En considérant de façon approximative que toutes les séquences de k éléments sont indépendantes les unes des autres, la probabilité qu'il n'y ait aucune séquence pure de longueur k dans un échantillon de N instances est environ $(1-P_{min}^{k})^{N}$. Donc $p_{k,N} \sim 1-(1-P_{min}^{k})^{N}. \sim N.P_{min}^{k}$. Par exemple, pour $P_{min} = 0,1$, k = 10 et $N = 10^{6}$, on a $p_{k,N} \sim 10^{-4}$, soit environ une chance sur 10000 de trouver une séquence pure de 10 modalités rares dans une suite de un million de valeurs distribuées aléatoirement. La contrainte d'effectif minimum de 50 par intervalle est dans ce cas très forte, et risque d'empêcher l'algorithme de discrétisation d'isoler des intervalles pertinents pour la discrétisation, ceci d'autant plus que la taille de l'ensemble d'apprentissage est petite.

Il faut alors trouver 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 discrétisation. Selon certains auteurs, l'effectif minimum théorique par cellule du tableau de contingence peut être ramené à 3, voire à 1 pour une seule classe en queue de distribution. On va alors relâcher la contrainte en en imposant un effectif minimum par intervalle de 6 (ce qui correspondrait à deux modalités cibles équidistribuées et un effectif minimum théorique de 3). Ce relâchement de la contrainte est agressif et donc risqué pour une évaluation fiable de la statistique du Khi2. On va renforcer cette contrainte en se basant sur le calcul précédent de la probabilité d'apparition d'une séquence pure dans un ensemble d'apprentissage. Pour une variable cible à deux modalités équidistribuées, la probabilité d'observer au moins une séquence pure de longueur k est $p_{k,N} \sim 1-(1-(1/2)^{k})^{N} \sim N(1/2)^{k}$. Pour une probabilité p donnée de ne pas discrétiser un attribut indépendant de l'attribut cible, il est donc légitime que l'effectif minimum soit au moins de la longueur k correspondante à $p=1-p_{k,N}$. Donc $N_{line} >= \log(1-p^{1/N})/\log(1/2)$. On peut vérifier que cette inégalité peut être approximée par $N_{line} >= \log_{2}(N)-\log_{2}(1-p)$, et donc que que l'effectif minimum par intervalle correspondant croît de façon logarithmique avec la taille de l'échantillon et avec le niveau de confiance demandé.

En résumé, on impose un effectif minimum dépendant de la taille de l'échantillon d'apprentissage 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
    - $N_{line} >= 6$
- Pour une probabilité p donnée, taille d'une séquence pure pour un attribut cible équidistribué
    - $N_{line} >= \log(1-p^{1/N})/\log(1/2)$

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

Le figure 4 trace d'une part l'effectif minimum théorique conservateur (dans l'hypothèse « raisonnable » d'un minimum de 3 par cellule) pour différents P_{min} et d'autre part la nouvelle valeur de N_{line} en fonction de la taille de l'ensemble d'apprentissage pour différentes valeurs du paramètre p. Pour $P_{min}=0,2$, l'effectif théorique minimum par ligne est 15. Dans ce cas, pour p = 0,95, cet effectif est atteint dès que le nombre d'instances est supérieur à 1000. Pour $P_{min}=0,1$, cet effectif minimum n'est pas respecté même pour de très grands nombres d'instances, mais la contrainte sera probablement respectée pour les autres modalités cibles.

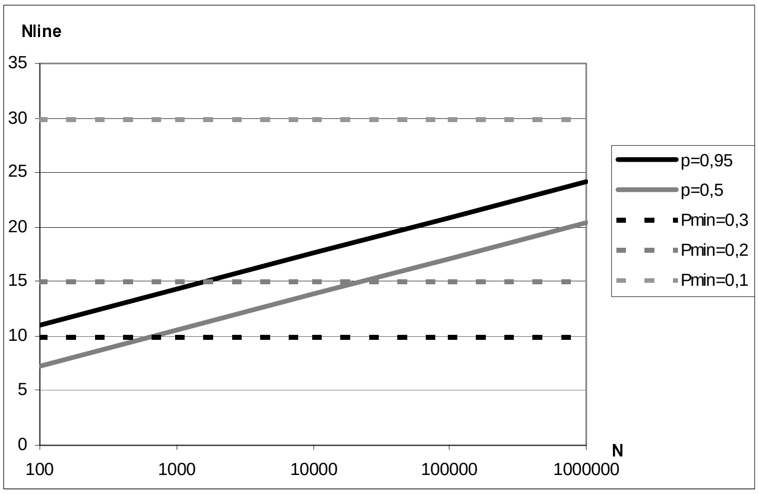


Figure 4 : Effectif minimum par ligne en fonction du nombre d'instances

Le compromis entre fiabilité du critère du Khi2 et finesse de discrétisation 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.

4.3.2 Post-optimisation des discrétisations

La méthode Khiops utilise une heuristique gloutonne d'optimisation de la discrétisation, en partant des intervalles élémentaires et en les agrégeant itérativement. Cette heuristique ascendante permet d'identifier des structures fines, mais la discrétisation finale constituée des structures principales peut avoir des frontières non optimales, si ces dernières ont été préalablement « capturées » lors des premières étapes de l'algorithme. Dans le cas limite d'une discrétisation finale en deux intervalles, il est possible que la frontière optimale ne soit pas atteinte, car préalablement agrégée dans un intervalle élémentaire au cours des étapes d'agrégation de l'algorithme. Pour un nombre d'intervalles donné, (Lechevallier, 1990) a montré que sous certaines conditions, il existait un algorithme de recherche de la partition optimal. Cet algorithme est de complexité algorithmique N^2 (N= nombre d'instances), ce qui le rend inutilisable sur les bases réelles utilisées en Data Mining. Il est cependant possible de proposer une heuristique simple, qui permet d'affiner les bornes d'une discrétisation. Le principe est de disposer d'un critère global d'évaluation de la discrétisation, et d'améliorer itérativement la discrétisation en déplaçant les bornes des intervalles. La méthode Khiops utilise un critère global avec des contraintes qui sont rattachées aux étapes de l'algorithme. On peut en fait extraire ce critère et ces contraintes sous la forme d'un problème d'optimisation global de la façon suivante :

- Minimiser la probabilité d'indépendance entre l'attribut discrétisé et l'attribut cible
- Sous les contraintes suivantes :
  - Chaque intervalle a un effectif minimum N_{min}
  - Chaque fusion d'intervalles adjacents entraîne une variation de la valeur du Khi2 global supérieure à MaxDeltaKhi2

Pour optimiser le point de coupure entre deux intervalles adjacents I_k et I_{k+1}, il suffit de parcourir tous les points de coupure potentiels dans les deux intervalles. On élimine alors les cas où les intervalles I_k et I_{k+1} ne respectent pas la contrainte d'effectif minimum, les cas où la variation du Khi2 d'une fusion entre I_{k-1} et I_k ou I_{k+1} et I_{k+2} ne respectent pas la contrainte du MaxDeltaKhi2, et on optimise le critère d'évaluation de la discrétisation sur les cas suivants.

Nous allons montrer que la recherche de la meilleure amélioration locale d'une discrétisation peut se faire de façon linéaire en fonction du nombre d'instances. En utilisant des techniques de bufferisation des calculs intermédiaires identiques à celles de l'algorithme principal Khiops, il est facile de vérifier que la recherche de la meilleure coupure a une complexité algorithmique linéaire vis à vis du nombre de points de coupure potentiels. Si l'on cherche la meilleure de toutes les coupures, tous les points de coupures potentiels auront été examinés pour chaque paire d'intervalles adjacents, donc au plus deux fois pour l'ensemble des valeurs de l'attribut. Une amélioration locale de la partition a donc une complexité linéaire en fonction du nombre d'instance.

On peut répéter cette étape de meilleure amélioration locale tant qu'il y a amélioration. Ce processus converge, mais on n'a pas de garantie sur le nombre d'étapes nécessaires. On peut alors borner le nombre d'étapes d'amélioration par log(N) pour se limiter à une complexité super-linéaire. En pratique, il est rare d'observer plus de deux ou trois étapes d'améliorations. Cette post-optimisation est néanmoins intéressante en permettant à peu de frais d'améliorer souvent la qualité des discrétisations obtenues. En particulier, dans le cas des discrétisation à deux intervalles, la partition optimale en deux intervalles est obtenue grâce à la post-optimisation.

5 Evaluation de l'algorithme Khiops robuste

5.1 Présentation

L'évaluation d'une méthode de discrétisation est un problème difficile, abordé de plusieurs façons dans la littérature. (Dougherty et al, 1995) comparent 4 méthodes de discrétisation en tant que prétraitement pour un prédicteur Bayesien Naïf et pour C4.5, et comparent le taux de bonne prédiction par cross-validation sur une quinzaine de jeux de données de la base UCI. (Liu et Al, 2002) comparent 8 méthodes de discrétisation en tant que prétraitement pour C4.5 sur une douzaine de jeux de données de l'UCI, et mesurent le temps de discrétisation, le temps de construction de l'arbre de décision avant et après discrétisation, le nombre de nœuds de l'arbre et le taux de bonne prédiction. (Zighed et al, 2001) comparent 7 méthodes de discrétisation en mesurant le taux de bonne prédiction des prédicteurs univariés pour les 21 attributs du jeu de données waveform. Cette dernière approche a le mérite d'isoler l'apport de la discrétisation, indépendamment de son emploi dans un prédicteur. Les conclusions générales sont que les différences de performances prédictives sont considérées comme minimes, mais que la méthode MDPLC (Fayyad, 1992) est communément reconnue comme étant une des méthodes les plus performantes.

Dans ce document, nous proposons une méthodologie d'évaluation des méthodes de discrétisation qui permet d'effectuer des comparaisons fiables sur plusieurs critères, et nous montrons que les différences observées sont significatives. Nous avons évalué six méthodes de discrétisation en utilisant quinze jeux de données standards extraits de la base UCI, comportant un nombre total de 181 attributs continus à discrétiser. Toutes 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 au seuil de confiance de 5% ont été évalués.

L'objet des expérimentations étant l'évaluation des méthodes de discrétisation, nous n'avons pris en compte que les attributs continus des jeux de données utilisés. L'expérimentation repose sur l'utilisation classique de la discrétisation en tant que prétraitement des données continues pour un prédicteur Bayesien Naïf, et surtout sur l'utilisation de prédicteurs univariés pour chaque attribut discrétisé afin d'évaluer les performance intrinsèques des méthodes de discrétisation sur un très grand nombre d'attributs. Les critères testés sont :

- Taux de bonne prédiction d'un prédicteur bayésien naïf
- Taux de bonne prédiction des prédicteurs univariés basés sur les discrétisations
- Robustesse des prédicteurs univariés : ratio du taux de bonne prédiction en test sur le taux de bonne prédiction en apprentissage
- Nombre d'intervalles des discrétisations
- Résistance au bruit : nombre d'intervalles lors de la discrétisation d'un attribut entièrement indépendant de l'attribut cible
- Complexité algorithmique

Nous détaillerons les procédures de tests utilisées lors de la présentation des résultats.

Les méthodes évaluées sont :

- Khiops : la méthode présentée dans ce document
- MDLPC : Minimum Description Length Principal Cut (Fayyad, 1992)
- ChiMerge : méthode ascendante basée sur le test du Khi2 (Kerber, 1991)
- ChiSplit : méthode descendante basée sur le test du Khi2
- EqualWidth : découpage en intervalle de largeur égale
- EqualFrequency : découpage en intervalles de fréquence égale

La méthode MDLPC est une méthode descendante qui découpe récursivement les intervalles en partant du domaine initial complet. Le critère MDLP évalue la quantité d'information contenue à la fois dans le modèle (la coupure) et les exceptions au modèle, et accepte la coupure si cette quantité d'information globale diminue après la coupure. Cette méthode ne nécessite aucun paramétrage.

La méthode ChiMerge est comme Khiops une méthode ascendante. Le critère utilisé est le critère du Khi2 appliqué localement à deux intervalles à fusionner. La fusion est acceptée si les deux intervalles sont suffisamment semblables pour un seuil de vraisemblance donné. Nous avons utilisé un seuil de 95% pour les expérimentations.

La méthode ChiSplit est une méthode descendante. Le critère utilisé est le critère du Khi2 appliqué localement aux deux sous-intervalles d'un intervalle à couper. La coupure est acceptée si les deux sous-intervalles sont suffisamment différents pour un seuil de vraisemblance donné. Nous avons utilisé un seuil de 95% pour les expérimentations.

Les méthodes EqualWidth et EqualFrequency sont des méthodes de discrétisation non supervisées, paramétrées par le nombre d'intervalles désiré. Nous avons adopté un découpage en 10 intervalles.

Les jeux de données extraits de la base UCI comportent tous au moins un attribut continu et au moins quelques dizaines d'instances par modalité cible. Ces jeux de données sont présentés dans le tableau 8, dont la dernière colonne représente le taux de bonne prédiction de la modalité cible majoritaire.

Tableau 8 : Jeux de données

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
German	24	0	1000	2	70,00
Heart	10	3	270	2	55,56
Hepatitis	6	13	155	2	79,35
Hypothyroid	7	18	3163	2	95,23
Ionosphere	34	0	351	2	64,10
Iris	4	0	150	3	33,33
Pima	8	0	768	2	65,10
SickEuthyroid	7	18	3163	2	90,74
Vehicle	18	0	846	4	25,77
Waveform	21	0	5000	3	33,92
Wine	13	0	178	3	39,89

5.2 Résultats par critère

5.2.1 Performance prédictive du prédicteur Bayesien Naïf

L’évaluation consiste ici à utiliser la discrétisation pour prétraiter les attributs continus avant leur utilisation par un prédicteur Bayesien Naïf. Le prédicteur Bayesien Naïf (Langley et al., 1992) prédit pour chaque instance la modalité cible la plus probable conditionnellement aux attributs descriptifs, en faisant l’hypothèse d’indépendance entre ces attributs descriptifs pour chaque modalité cible. Après l’étape de discrétisation, les probabilités des attributs continus sont estimées par comptage dans chaque intervalle.

L’évaluation du prédicteur Bayesien Naïf est menée par cross-validation sur les quinze jeux de données de l’UCI. Le tableau 9 montre la moyenne et l’écart type des taux de bonne prédiction pour chacun des jeux de données. Les gains significatifs pour la méthode Khiops sont marqués par des **+**, et les pertes significatives par des **-**.

Tableau 9 : Taux de bonne prédiction du prédicteur Bayesien Naïf avec différentes méthodes de discrétisation

Dataset	Khiops	MDLPC	ChiMerge	ChiSplit	Eq. Width	Eq. Freq.
Adult	84,2 ±0,5	84,4 ±0,5	77,8 ±0,7 **+**	84,3 ±0,5	81,2 ±0,4 **+**	81,1 ±0,6 **+**
Australian	77,1 ±3,4	77,4 ±3,6	75,1 ±4,6	78,1 ±3,5	71,0 ±5,3 **+**	80,4 ±2,4 **-**
Breast	97,1 ±1,4	97,1 ±1,1	90,1 ±3,5 **+**	97,0 ±1,7	96,6 ±1,7	97,4 ±1,4
Crx	76,7 ±6,0	76,5 ±5,8	71,4 ±5,9 **+**	78,0 ±7,0	70,3 ±3,6 **+**	79,7 ±6,1 **-**
German	71,4 ±3,6	72,5 ±1,8	74,3 ±3,6 **-**	75,6 ±4,2 **-**	75,5 ±3,6 **-**	75,5 ±3,9 **-**
Heart	81,5 ±7,0	80,7 ±8,6	72,2 ±6,3 **+**	78,9 ±8,8	81,1 ±6,5	80,7 ±6,6
Hepatitis	80,2 ±9,6	76,8 ±13,	81,5 ±11,	78,9 ±9,5	82,6 ±8,5	78,8 ±11,
Hypothyroid	98,6 ±0,9	98,7 ±0,6	98,2 ±0,6 **+**	98,5 ±1,0	97,4 ±0,6 **+**	97,6 ±0,9 **+**
Ionosphere	89,2 ±3,5	90,9 ±4,4 **-**	86,1 ±5,2	86,3 ±4,0	89,5 ±4,4	91,2 ±3,9
Iris	94,0 ±3,6	92,7 ±2,0	94,7 ±2,7	94,0 ±3,6	95,3 ±4,3	94,7 ±5,0
Pima	74,7 ±2,5	76,2 ±2,1 **-**	72,0 ±2,7 **+**	75,0 ±3,0	74,7 ±3,1	74,0 ±3,5
SickEuthyroid	95,9 ±1,1	95,9 ±1,1	96,3 ±1,3	95,8 ±1,0	92,9 ±1,8 **+**	93,2 ±1,1 **+**
Vehicle	61,2 ±1,9	60,2 ±2,3	64,5 ±3,8 **-**	63,1 ±4,1	63,6 ±3,3 **-**	61,4 ±3,3
Waveform	81,0 ±1,2	80,8 ±0,8	75,9 ±1,6 **+**	80,0 ±1,8	80,8 ±1,2	80,7 ±1,2
Wine	95,6 ±2,2	96,7 ±3,7	96,6 ±4,5	95,0 ±3,9	95,5 ±6,5	96,6 ±3,7
Mean	83,9	83,8	81,8	83,9	83,2	84,2
+ number		0	7	0	5	3
- number		2	2	1	2	3

La comparaison fine entre Khiops et MDLPC montre que pour 15 jeux de données, Khiops gagne 8 fois et perd 7 fois (dont 2 fois de façon significative). La situation est encore plus incertaine dans la comparaison avec EqualFrequency, où Khiops gagne 7 fois (dont 3 de façon significative) et perd 8 fois (dont 3 de façon significative). Les méthodes Khiops, MDLPC, ChiSplit et EqualFrequency obtiennent donc des résultats comparables, meilleurs que ceux obtenus par les méthodes ChiMerge et EqualWidth. Ces résultats permettent de distinguer deux groupes de méthodes, mais ne sont pas décisifs pour classer plus avant les méthodes.

5.2.2 Performance prédictive intrinsèque

Afin d’analyser plus en détail les performances des méthodes de discrétisation, nous avons procédé à de nouvelles expérimentations en mesurant la performance prédictive des discrétisations pour chacun des attributs continus des 15 jeux de données précédemment utilisés. Cela correspond à l’utilisation de 181 jeux de données mono-attributs, ce qui d’une part apporte une fiabilité statistique aux résultats obtenus, et d’autre part permet d’évaluer la performance prédictive intrinsèque des méthodes de discrétisation, sans le biais dû à l’utilisation d’un prédicteur agrégé (ici le prédicteur Bayesien Naïf).

La table de résultats est trop volumineuse pour être reproduite dans ce document. Elle est résumée dans le tableau 10, qui indique pour chaque jeu de données la moyenne des taux de bonne prédiction par attribut et le nombre de gains et de pertes significatives dans les comparaisons avec Khiops.

Tableau 10 : Moyenne des taux de bonne prédiction, nombre de gains et pertes par jeu de données, pour les prédicteurs élémentaires univariés

Dataset	Khiops	MDLPC			ChiMerge			ChiSplit			Eq. Width			Eq. Freq.		
			+	-		+	-		+	-		+	-		+	-
Adult	77,3	77,3	0	2	75,7	2	2	77,3	0	2	76,8	2	1	76,6	2	1
Australian	64,8	65,0	0	0	64,7	0	0	65,1	0	0	61,4	3	0	65,7	0	0
Breast	85,8	86,1	0	1	85,6	0	1	85,9	0	1	86,0	0	1	85,7	1	1
Crx	65,0	65,2	0	0	63,8	2	0	65,3	0	0	61,1	3	0	65,6	0	1
German	70,1	70,0	0	0	70,0	0	0	70,1	0	0	70,1	0	2	70,0	0	0
Heart	64,4	64,0	0	0	64,0	0	0	63,8	0	0	63,9	2	0	64,5	1	0
Hepatitis	79,6	79,3	0	0	77,8	3	0	79,3	0	0	79,8	0	0	79,9	0	0
Hypothyroid	96,1	96,1	0	1	96,0	3	0	96,1	1	0	95,4	3	1	95,2	3	1
Ionosphere	79,7	77,6	10	2	75,7	21	0	79,5	4	3	73,9	19	1	75,0	22	0
Iris	78,8	75,5	1	0	77,0	0	0	78,8	0	0	76,5	1	0	76,3	0	0
Pima	66,3	66,1	0	0	65,6	2	0	66,5	0	0	66,8	0	1	66,3	0	1
SickEuthyroid	91,3	91,3	0	0	91,3	1	0	91,3	0	0	90,7	2	0	91,0	1	0
Vehicle	41,5	40,5	4	0	41,4	2	1	42,1	0	3	40,8	3	0	40,3	3	0
Waveform	49,3	49,3	0	0	48,7	6	0	49,1	4	0	49,2	3	3	49,5	1	4
Wine	60,0	60,1	0	1	59,6	1	0	60,4	0	1	61,4	2	2	60,8	1	2
Synthesis	68,6	68,0	15	7	67,4	43	4	68,6	9	10	67,2	43	12	67,6	35	11

L'expérimentation basée sur l'analyse univariée des méthodes de discrétisation est plus riche d'enseignements que l'expérimentation précédente basée sur le prédicteur Bayesien Naïf. Les résultats sont plus contrastés, tant en nombre de gains ou de pertes significatives qu'en valeur moyenne synthétique de tous les résultats obtenus sur l'ensemble des 181 attributs. L'analyse des résultats montre que les méthodes supervisées (sauf ChiMerge) obtiennent des résultats nettement meilleurs que les méthodes non supervisées. La méthode ChiMerge est légèrement meilleure que la méthode EqualWidth, mais moins performante que la méthode Equal Fréquency. La méthode MDLPC est nettement meilleure que la méthode EqualFrequency. Les méthodes Khiops et ChiSplit se détachent en tête de façon indiscernable, en surpassant la méthode MDLPC.

Afin de mieux comprendre l'importance des différences entre les méthodes, on va s'attacher à comparer en détail les résultats des méthodes Khiops et MDLPC. L'écart entre les valeurs moyennes synthétiques (68,6 pour Khiops et 68,0 pour MDLPC) peut paraître faible (0,6%), mais il faut le comparer à la valeur moyenne du prédicteur majoritaire qui est de 57,4. L'amélioration relative par rapport au prédicteur majoritaire est alors supérieure à 5%, ce qui n'est pas négligeable. D'autre part, cet écart absolu de 0,6% est une moyenne sur une grande variété d'attributs, et il est intéressant d'examiner en détail comment se répartissent les différences de taux de bonnes prédictions sur l'ensemble des 181 attributs ayant servi à l'évaluation. La figure 5 représente les fonctions de répartition des différences de performance entre Khiops et toutes les autres méthodes.

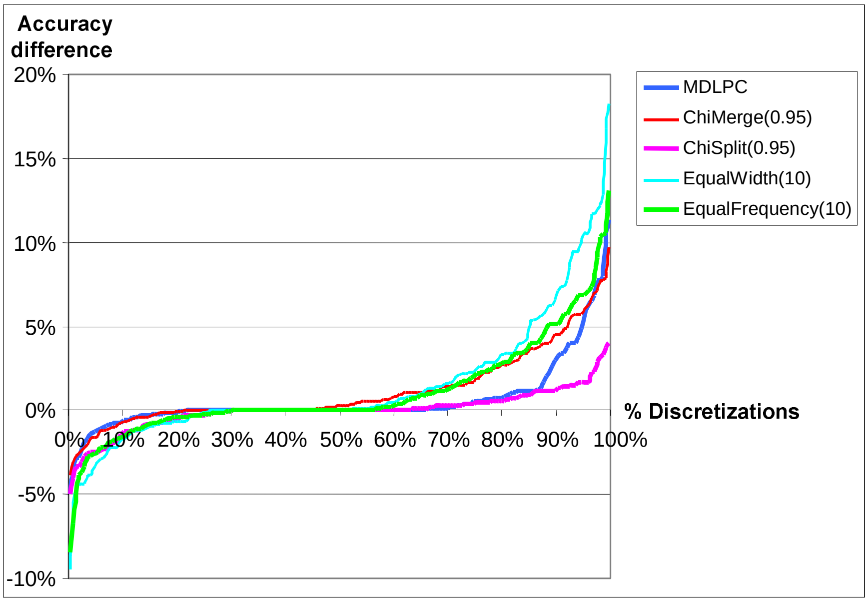


Figure 5 : Répartition des différences de taux de bonne prédiction entre Khiops et les autres méthodes

La partie gauche de la figure représente les cas où Khiops est moins performant que les autres méthodes, et la partie droite les cas où Khiops est meilleur. Dans 40% des cas environ (entre les abscisses 20% et 60%), toutes les méthodes sont équivalentes. Cela correspond à des attributs ayant une seule valeur ou très peu de valeurs, qui ne sont pas utilisables en fait pour comparer les discrétisations. Si l'on compare Khiops à MDLPC, Khiops est de 0 à 3% moins bon que MDLPC dans environ 10% des cas de discrétisation, mais est de 3 à 10% meilleur dans environ 10% des cas. On voit alors clairement que la différence moyenne de 0,6% entre Khiops et MDLPC est significative, et qu'elle se traduit en fait par des différences beaucoup plus importantes.

On peut dès lors considérer que les différences observées entre les méthodes sont significatives, et que la valeur moyenne synthétique (sur 181 attributs discrétisés chacun 10 fois) est un bon indicateur global permettant de comparer les performances des méthodes.

Sur le critère du taux de bonne prédiction, on peut donc classer les méthodes de la façon suivante :

1. Khiops, ChiSplit
2. MDLPC
3. EqualFrequency, ChiMerge
4. EqualWidth

5.2.3 Robustesse

On a cette fois mesuré la robustesse des discrétisations en prenant le même protocole de test sur l'ensemble de tous les attributs continus des jeux de données. La robustesse mesure la dégradation de performance prédictive entre la phase d'apprentissage et de test. On a ici mesuré le ratio du taux de bonne prédiction en test sur le taux de bonne prédiction en apprentissage. Ce ratio est en principe borné par 100%. Les méthodes ayant tendance à sur-apprendre ont une robustesse plus faible. Le tableau 11 résume les résultats obtenus lors de l'évaluation de la robustesse sur l'ensemble des attributs des jeux de données.

Tableau 11 : Moyenne des taux de robustesse, nombre de gains et pertes par jeu de données, pour les prédicteurs élémentaires univariés

Dataset	Khiops	MDLPC			ChiMerge			ChiSplit			Eq. Width			Eq. Freq.		
			+	-		+	-		+	-		+	-		+	-
Adult	100	100	0	1	95,3	2		99,9	1	0	100	0	1	100	1	1
Australian	97,8	98,1	0	0	93,5	4		97,2	1	0	100	0	1	99,0	0	0
Breast	99,6	100	0	1	97,2	2		99,0	1	0	99,7	0	0	99,6	0	1
Crx	97,9	98,4	0	0	92,1	4		97,4	1	0	99,4	0	1	99,1	0	4
German	99,9	99,9	0	0	99,8	0		99,9	0	0	99,9	0	0	99,9	0	1
Heart	99,6	99,2	0	0	95,2	4		97,1	2	0	98,0	1	0	97,8	1	0
Hepatitis	98,5	99,5	0	0	92,8	3		97,8	0	0	98,2	1	0	99,0	0	0
Hypothyroid	99,9	100	0	1	99,7	3		99,8	1	0	100	0	1	100	0	1
Ionosphere	97,9	98,0	1	0	87,5	32		96,4	9	0	99,4	0	2	98,7	0	1
Iris	97,8	94,5	1	0	94,4	0		97,5	0	0	96,3	0	0	95,9	0	0
Pima	99,3	99,1	0	0	94,8	5		98,2	1	0	98,6	1	0	98,5	0	0
SickEuthyroid	100	100	0	0	99,8	2		99,9	0	0	100	1	0	100	0	0
Vehicle	96,4	96,1	1	1	92,0	7		94,5	4	0	95,4	1	0	93,7	4	0
Waveform	98,7	98,7	0	0	92,6	21		96,8	12	0	98,7	1	2	98,4	3	3
Wine	94,7	96,1	0	3	87,4	5		90,7	4	0	95,1	1	2	92,4	2	1
Synthesis	98,4	98,5	3	7	93,4	94		97,1	37	0	98,6	7	10	98,0	11	13

Sans surprise, les méthodes non supervisées EqualWidth et EqualFrequency, qui ne font aucune hypothèse sur les données, ont une très bonne robustesse. Le fait de ne pas atteindre 100% est dû au découpage en un nombre d'intervalles assez important (10), et en la fluctuation statistique des distributions entre ensemble d'apprentissage et de test. De plus, la procédure de validation croisée en 10 étapes associée à des ensemble parfois petits (moins de 200 instances) fait qu'une seule erreur dans l'ensemble de test (moins de 20 instances) se traduit par une perte de robustesse d'environ 5%. Quoiqu'il en soit, les méthodes EqualWidth et EqualFrequency ont une très bonne robustesse moyenne, qui peut servir de référence. Les méthodes supervisées Khiops et MDLPC ont une excellente robustesse, équivalentes à celles des méthodes non supervisées. La méthode ChiSplit est nettement moins robuste que la méthode Khiops. Elle est nettement moins robuste que Khiops, en obtenant 37 fois une robustesse significative inférieure. La méthode ChiMerge, extrêmement loin de toutes les autres méthodes, est manifestement sujette au sur-apprentissage.

Sur le critère de la robustesse, on peut donc classer les méthodes de la façon suivante :

1. Khiops, MDLPC, EqualFrequency, EqualWidth
2. ChiSplit
3. ChiMerge

5.2.4 *Taille des discrétisations*

La taille des discrétisations est simplement le nombre d'intervalles obtenus sur l'ensemble d'apprentissage par les différentes méthodes de discrétisation. Le tableau 12 résume les résultats obtenus lors de l'évaluation de la taille des discrétisations sur l'ensemble des attributs des jeux de données.

Tableau 12 : Moyenne des nombres d'intervalles, nombre de gains et pertes par jeu de données, pour les prédicteurs élémentaires univariés

Dataset	Khiops			MDLPC			ChiMerge			ChiSplit			Eq. Width			Eq. Freq.		
					+	-		+	-		+	-		+	-		+	-
Adult	8,5	0		8,8	2	3	1264	0	7	28,2	0	6	9,4	2	4	6,6	4	2
Australian	2,1	0		2,0	0	0	16,1	0	6	5,2	0	6	8,1	0	6	8,8	0	6
Breast	2,6	0		2,9	0	5	11,6	0	9	4,9	0	9	9,2	0	10	5,9	0	10
Crx	2,1	0		2,1	0	0	15,8	0	6	5,1	0	6	8,2	0	6	8,7	0	6
German	1,3	0		1,2	2	0	2,4	0	12	2,0	0	12	3,8	0	23	3,4	0	21
Heart	1,7	0		1,7	0	0	5,0	0	5	2,5	0	5	5,9	0	8	6,1	0	8
Hepatitis	1,7	0		1,4	1	0	6,4	0	6	2,8	0	5	8,6	0	6	9,2	0	6
Hypothyroid	3,5	0		3,1	3	0	15,3	0	7	6,0	0	7	9,6	0	7	8,3	0	7
Ionosphere	4,3	0		3,9	11	6	30,0	0	32	8,0	0	30	9,4	0	32	8,9	0	32
Iris	2,8	0		2,8	1	0	3,7	0	3	3,6	0	3	9,7	0	4	9,5	0	4
Pima	2,3	0		2,1	2	1	13,2	0	8	5,0	0	8	9,5	0	8	9,3	0	8
SickEuthyroid	3,4	0		3,0	3	0	17,2	0	6	5,8	0	5	9,6	0	7	8,3	0	6
Vehicle	4,0	0		3,9	3	3	9,7	0	18	8,0	0	18	9,6	0	18	9,6	0	18
Waveform	4,5	0		4,9	1	9	49,0	0	21	13,6	0	21	10,0	0	21	10,0	0	21
Wine	2,6	0		2,8	1	3	6,7	0	13	4,8	0	12	9,8	0	13	10,0	0	13
Synthesis	3,3	0		3,2	30	30	66,1	0	159	7,2	0	153	8,5	2	173	8,0	4	168

Cette mesure est sans intérêt pour les méthodes non supervisées, dont on a paramétré le nombre d'intervalles. On remarque que ces méthodes aboutissent à un peu moins que les 10 intervalles demandés, à cause des attributs pour lesquels il n'y a pas assez de valeurs différentes pour trouver un découpage effectif en 10 intervalles. Les méthodes Khiops et MDLPC produisent les discrétisations les plus petites, et sont indiscernables pour ce critère : il y a exactement autant de différences significatives dans un sens et dans l'autre quand on compare ces deux méthodes, et la taille moyenne globale est quasiment identique. Le nombre important de différences significatives observées (60) est un peu artificiel, et provient du fait que la valeur mesurée est le nombre d'intervalles (qui ne prend que des valeurs entières). La méthode ChiSplit produit des discrétisation plus de deux fois volumineuses en nombre d'intervalles. La méthode ChiMerge génère des nombres d'intervalles extrêmement importants. Il est à noter que pour la base Adult, ce nombre d'intervalles est considérable (plus de 1000 intervalles en moyenne pour 50000 individus). On remarque que cette base nécessite apparemment un nombre d'intervalles assez important, y compris pour les méthodes les plus performantes. C'est la seule base pour laquelle Khiops ou MDLPC produisent plus d'intervalles que EqualFrequency (avec un paramètre de 10 intervalles demandés).

Sur le critère de la taille des discrétisations, on peut donc classer les méthodes supervisées de la façon suivante :

1. Khiops, MDLPC
2. ChiSplit
3. ChiMerge

5.2.5 *Résistance au bruit*

Dans cette expérimentation, on a évalué le comportement des différentes méthodes en présence d'un attribut continu indépendant de l'attribut cible. Pour cela, on a utilisé des jeux de données artificiels de taille 100, 1000, 10000 et 100000, comportant un attribut continu descriptif ayant toutes ses valeurs distinctes, et un attribut cible à deux modalités équiréparties, choisies de façon aléatoire. Pour chaque méthode, on a mesuré le pourcentage des discrétisations aboutissant à un seul intervalle, et la taille des discrétisation quand elles aboutissaient à plusieurs intervalles. L'expérimentation a été effectuée 1000 fois pour obtenir un résultat fiable. L'expérimentation ne concerne que les méthodes de discrétisation supervisées (pour lequel on ne paramètre pas le nombre d'intervalles à obtenir). Les figures 6 et 7 présentent les résultats de l'expérimentation obtenus par les différentes méthodes, pour différents paramétrages.

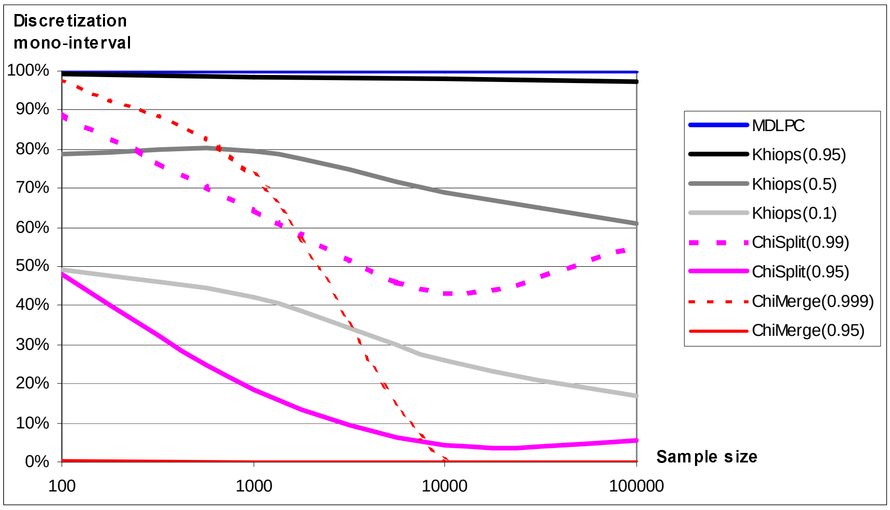


Figure 6 : Pourcentage de discrétisation mono-intervalles en présence d'un attribut indépendant de l'attribut cible

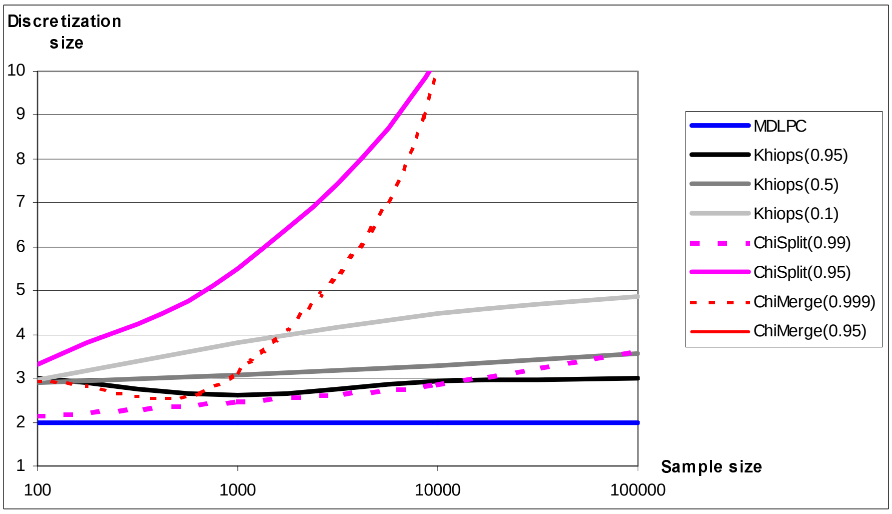


Figure 7 : Taille des discrétisations multi-intervalles en présence d'un attribut indépendant de l'attribut cible

MDLPC est très résistante au bruit, et aboutit presque toujours à un seul intervalle. Des expérimentations plus poussées réalisées 100000 fois ont montré que le pourcentage de discrétisation mono-intervalles est de l'ordre 99,95% et qu'il croît avec le nombre d'instances (99,90% pour 100 instances, 99,94% pour 1000 instances et 99,98% pour 10000 instances). Dans les rares cas de discrétisations multi-intervalles, MDLPC aboutit systématiquement à deux intervalles.

ChiMerge avec son paramètre par défaut à 0,95 aboutit systématique à plusieurs intervalles, et le nombre d'intervalles croît linéairement avec la taille du jeu de données. Ce nombre, d'environ 10% de la taille de l'échantillon n'est pas visualisable avec l'échelle choisie sur la figure 7. Il faut utiliser un paramètre de 0,999 pour obtenir une différence de comportement significative. Dans ce cas, on passe de presque 100% de discrétisation mono-intervalles pour les jeux de données de taille 100 à quasiment 0% de discrétisations mono-intervalles pour des jeux de données de taille 10000. La taille des discrétisations croît également très vite avec la taille des jeux de données.

ChiSplit avec son paramètre par défaut à 0,95 aboutit à un seul intervalle dans environ la moitié des cas pour des jeux de données de taille 100, mais ce taux tombe à 5% pour des jeux de données de taille 10000. L'utilisation d'un paramètre à 0,99 permet d'obtenir un comportement stable pour la méthode, correspond à un taux de discrétisations mono-intervalles de l'ordre de 60%.

Khiops a un comportement relativement stable avec la taille de l'échantillon. Son paramétrage, qui permet en théorie de contrôler le pourcentage des discrétisations mono-intervalles dans le cas de l'expérimentation, s'avère comme prévu pessimiste. Ainsi, on aboutit à environ 98% de discrétisations mono-intervalles pour le paramètre 0,95, à entre 60% et 80% pour le paramètre 0,50 et à entre 20% et 50% pour le paramètre 0,10. Ceci est conforme aux courbes de la figure 2, où l'on voit que la modélisation théorique du comportement de l'algorithme se comporte comme une borne inférieure pour le comportement observé. La taille des discrétisations multi-intervalles est d'environ 3 pour le paramètre 0,95 et croît légèrement quand la valeur du paramètre diminue.

De façon générale, on observe que les méthodes ChiMerge et ChiSplit sont très sensibles à la taille des jeux de données, le taux de discrétisations mono-intervalles et la taille des discrétisations augmentent rapidement avec la taille des jeux de données. Pour des critères statistiques équivalents, l'approche ascendante utilisée dans ChiMerge nécessite un paramétrage plus strict que l'approche descendante utilisée dans ChiSplit. MDLPC et Khiops ont un comportement stable indépendamment de la taille du jeu de données. MDLPC aboutit quasiment systématiquement à un seul intervalle alors que Khiops contrôle ce comportement par son paramétrage. Il est notable que les discrétisations multi-intervalles de MDLPC comportent 2 intervalles (un seul split a été effectué), alors qu'elles comportent environ 3 intervalles pour Khiops (un pattern émergeant du « bruit » a été isolé au milieu de deux intervalles standard). Ce comportement correspond à la différence d'approche descendante ou ascendante utilisée par les deux méthodes.

On peut remarquer le côté paradoxal de cette expérimentation. Il paraît évident que les méthodes de discrétisation supervisées sont d'autant meilleures qu'elles génèrent peu d'intervalles en présence d'un attribut indépendant de l'attribut cible. Pourtant, une méthode produisant systématiquement un seul intervalle n'aurait aucun intérêt pour la prédiction. Inversement, une méthode produisant de nombreux intervalles n'est pas pénalisée quand on mesure son taux de bonne prédiction. Sur chaque intervalle, la prédiction est en fait très proche de celle du prédicteur majoritaire (optimale dans le cas de l'expérimentation), et les petites fluctuations de bonne prédiction entre les intervalles ont tendance à se compenser globalement. Ce critère est alors très intéressant pour éclairer la stratégie d'apprentissage des différentes méthodes Les comportements observés vont de MDPLC qui produit presque toujours un seul intervalle à ChiMerge et ChiSplit qui en produisent d'autant plus que la taille du jeu de données augmente, en passant par Khiops qui contrôle la probabilité de produire des intervalles.

5.2.6 Complexité algorithmique

La méthode la plus simple, EqualWidth, est la seule qui soit linéaire en fonction de la taille N de l'échantillon. Toutes les autres méthodes (y compris EqualFrequency) nécessitent un tri des valeurs de l'attribut à discrétiser, et sont donc au moins en N.log(N). Les méthodes de discrétisation supervisée descendantes (MDLPC et ChiSplit) sont assez simples à implémenter : il suffit de savoir chercher le meilleur point de coupure d'un intervalle, puis de réitérer la procédure récursivement. Si l'arbre des coupures obtenues est assez équilibré, on obtient une complexité en N.log(N). Cependant, à ma connaissance, il n'y a pas de telle garantie d'équilibre de l'arbre, ce qui pourrait théoriquement entraîner une complexité algorithmique en N^2. Les méthodes de discrétisation supervisée ascendante (Khiops et ChiMerge) sont plus compliquées à implémenter si l'on veut garantir une complexité algorithmique en N.log(N). Elles nécessitent des structures algorithmiques permettant de bufferiser les résultats des calculs intermédiaires et de mémoriser la liste triée des meilleures fusions d'intervalles dans un arbre binaire de recherche équilibré.

En pratique, il est difficile de mesurer et comparer précisément les temps de discrétisation. Pour des implémentations optimisées, on obtient en effet des temps de discrétisation de l'ordre de une seconde par attribut pour des jeux de données de taille 100000. L'utilisation de très grands jeux de données permet d'observer des petites différences de l'ordre de 10 à 20% entre les méthodes, ce qui est peu significatif en regard des autres critères étudiés. En fait, le temps dévolu au tri des valeurs est largement dominant dans le temps de discrétisation. En effet, après tri des valeurs, on peut regrouper les instances par valeurs, ce qui permet de se ramener à un nombre V de valeurs en pratique largement inférieur au nombre d'instance. Le temps de discrétisation (hors tri) est alors en V.log(V), petit devant N.log(N), ce qui ne permet pas de différencier les méthodes (hormis EqualWidth) sur ce critère.

5.3 Analyse multi-critères des résultats

La comparaison des méthodes à l'issue des expérimentations permet de comparer et de classer les méthodes sur chacun des critères étudiés. Il est intéressant de procéder à une analyse multi-critères des résultats, afin de mieux comprendre les liens entre ces critères, notamment pour la performance prédictive, la robustesse et la taille des discrétisations.

Rappelons dans un premier temps quelques notions de l'analyse multi-critères. On dit qu'une solution **domine** (ou est **non inférieure** à) une autre solution si elle est meilleure sur tous les critères. Une solution ne peut être dominée si toute amélioration sur un des critères entraîne une détérioration sur un autre critère. Une telle solution est un **optimum de Pareto**. La **surface de Pareto** (**courbe de Pareto** pour 2 critères) est l'ensemble de tous les optima de Pareto.

5.3.1 Présentation des expérimentations

On a repris ici les résultats des discrétisations sur les attributs élémentaires des jeux de données et utilisé la valeur moyenne (sur les 1810 discrétisations effectuées) comme indicateur de performance pour le taux de bonne prédiction, la robustesse et le nombre d'intervalles des méthodes de discrétisation.

On a inclus les versions précédentes de Khiops pour mesurer l'apport des évolutions. La version KhiopsBasic est la version initiale avec effectif minimum par intervalle en racine carrée de la taille du jeu de données. La version KhiopsRobust est la version avec contrôle statistique du sur-apprentissage. La version Khiops finale est identique à la version KhiopsRobust, avec en plus la post-optimisation des points de coupure des intervalles.

Afin d'étudier l'influence du paramétrage, on a refait les expérimentations pour les valeurs suivantes de paramètre des méthodes :

- Khiops : 1e-9 ; 1e-8 ; … ; 1e-4 ; 0,001 ; 0,01 ; 0,1 ; 0,5 ; 0,9 ; 0,95 ; 0,99, 0,999 ; 0,99999 ; 0,99999
- KhiopsRobust : 0,001 ; 0,01 ; 0,1 ; 0,5 ; 0,9 ; 0,95 ; 0,99, 0,999
- KhiopsBasic et MDLPC : pas de paramètre
- ChiMerge et ChiSplit : 0,5 ; 0,9 ; 0,99 ; 0,95 ; 0,999 ; 0,99999 ; 0,99999
- EqualFrequency et EqualWidth : 1 ; 2 ; 3 ; … ;18 ; 19 ; 20

Les positions des points sont des évaluations statistiques des critères étudiés, et doivent être interprétés avec précaution dans le cadre de l'analyse multi-critères. En particulier, on dira qu'une solution domine strictement une autre solution pour un critère donné si les différences observées sont significatives. Afin d'étalonner les différences de valeurs significatives, nous avons effectué le test de Student pour des paires de points représentatifs. Nous disposons de 181 observations, représentant pour chaque attribut de chaque jeu de données la valeur moyenne d'un critère de discrétisation mesurée à l'issue d'une validation croisée en 10 étapes. On donne dans le tableau 13 les résultats de comparaisons par paires les plus pertinents, pour le critère du taux de bonne prédiction.

Tableau 13 : Différences significatives de taux de bonne prédiction pour quelques paires de méthodes représentatives

Méthode 1	Accuracy	Méthode 2	Accuracy	Diff	Student t-value	Prob significatif
Khiops(0,95)	68,57%	MDLPC	68,00%	0,57%	3,71	99,9%
Khiops(0,95)	68,57%	KhiopsBasic	68,38%	0,19%	1,67	90%
Khiops(0,95)	68,57%	KhiopsRobust(0,95)	68,46%	0,11%	2,47	98%
Khiops(0,5)	68,71%	Khiops(0,95)	68,57%	0,14%	1,99	95%
Khiops(0,5)	68,71%	KhiopsBasic	68,38%	0,33%	3,48	99,9%
Khiops(0,5)	68,71%	KhiopsRobust(0,5)	68,66%	0,05%	1,40	83%
Khiops(0,5)	68,71%	ChiMerge(0,999)	68,43%	0,28%	3,52	99,9%
Khiops(0,5)	68,71%	ChiSplit(0,95)	68,60%	0,11%	1,59	88%
Khiops(0,5)	68,71%	ChiSplit(0,99)	68,67%	0,04%	0,60	45%
Khiops(0,5)	68,71%	ChiSplit(0,999)	68,55%	0,16%	2,08	96%

On voit ainsi qu'une différence d'environ 0,15% de taux de bonne prédiction est significative à 95%. Ainsi, la meilleure valeur de Khiops (Khiops(0,5)) domine significativement les méthodes MDLPC, ChiMerge et KhiopsBasic, mais ne se détache pas significativement de la meilleure valeur de ChiSplit (ChiSplit(0,99)).

Pour le critère de la robustesse, une analyse similaire montre qu'une différence d'environ 0,25% est significative à 95%. Pour le critère du nombre d'intervalles, une différence d'environ 0,25 intervalles est significative à 95%. Dans les diagrammes à deux dimensions utilisés pour présenter les résultats pour chaque couple critère, la graduation secondaire sur l'intérieur des axes reprend ces valeurs pour rappeler la taille de la « maille » significative permettant de différencier significativement les performances des méthodes.

5.3.2 *Performance prédictive et robustesse*

La figure 8 présente l'ensemble des résultats sur le plan des critères de taux de bonne prédiction et de robustesse.

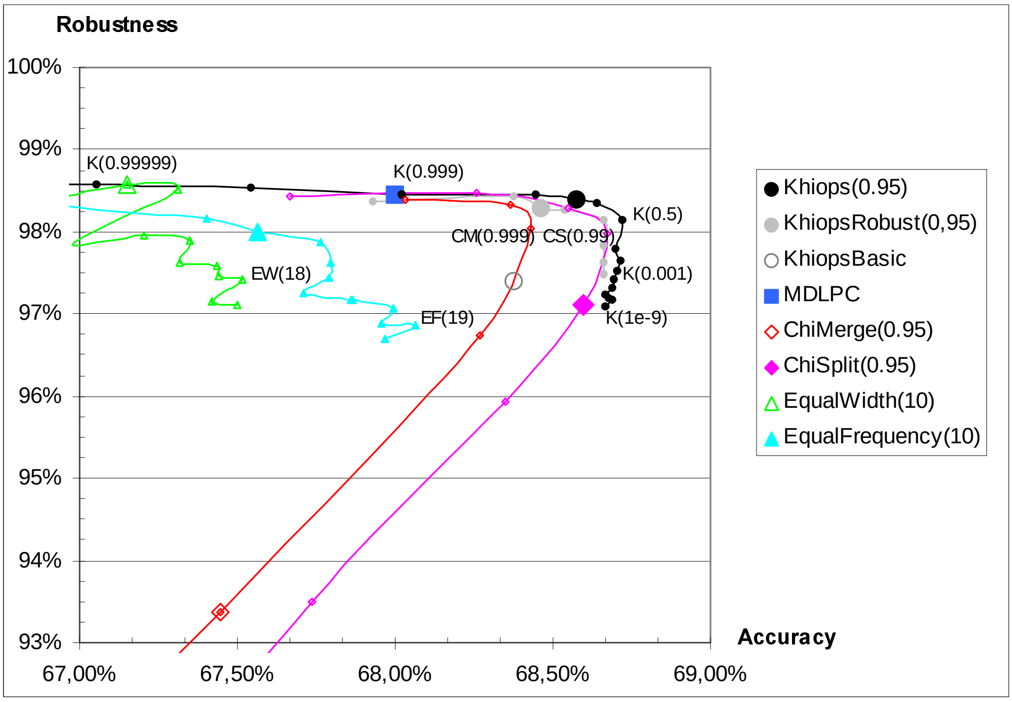


Figure 8 : Evaluation des méthodes de discrétisation pour les critères de taux de bonne prédiction et de robustesse

Les méthodes non supervisées EqualWidth et EqualFrequency sont très largement dominées par l'ensemble des autres méthodes, la méthode EqualFrequency étant significativement meilleure que EqualWidth sur le critère du taux de bonne prédiction. Par ailleurs, ces méthodes sont relativement instables vis à vis de leur paramétrage. Elles améliorent leur taux de bonne prédiction en dégradant leur robustesse quand le nombre d'intervalles demandé augmente.

Conformément aux premières expérimentations, la méthode ChiMerge est la moins performante des méthodes supervisées testées. Pour une petite plage de paramètres entre 0,999 et 0,9999, elle arrive à dépasser MDLPC en taux de bonne prédiction. Néanmoins, ChiMerge est extrêmement sensible à son paramétrage, de très petites variations de paramètre entraînant des variations très importantes de performances. Trouver la bonne plage de valeurs du paramétrage par jeux de données, voire par attribut la rend non compétitive face à MDLPC qui ne nécessite aucun paramètrage.

La méthode MDLPC obtient de très bons résultats en termes de robustesse, mais est dépassée par les autres méthodes, notamment ChiSplit et Khiops sur des plages importantes de leur paramétrage.

ChiSplit obtient de très bonnes performances. Pour un paramètre de 0,99, elle obtient presque les meilleures performances en taux de bonne prédiction. Pour un paramètre de 0,999, elle a des performances similaires à celle de Khiops(0,95) sur les deux critères. Néanmoins, bien que plus stable que ChiMerge, ChiSplit est également très sensible à la valeur de son paramétrage.

Il est intéressant de remarquer que les deux courbes pour ChiMerge et ChiSplit ont des formes similaires. Un paramétrage très stricte (proche de 1) entraîne une robustesse très importante, mais une performance prédictive assez faible : le paramétrage est trop « conservateur » et l'apprentissage est modéré. Quand le paramétrage se relâche, l'apprentissage devient effectif, ce qui se traduit par une augmentation du taux de bonne prédiction avec une dégradation négligeable de la robustesse. On obtient alors un maximum (paramètre de 0,999 pour ChiMerge et 0,99 pour ChiSplit dans cette expérimentation), au delà duquel la méthode est clairement sujette à sur-apprentissage, ce qui dégrade rapidement à la fois la robustesse et le taux de bonne prédiction.

La méthode Khiops obtient les meilleurs résultats sur l'ensemble des deux critères, et sa courbe semble située sur la courbe optimale de Pareto pour l'ensemble des méthodes testées. Pour une robustesse donnée, aucune méthode ne dépasse Khiops en taux de bonne prédiction, et pour un taux de bonne prédiction donné, aucune méthode ne dépasse Khiops en robustesse. Par exemple, pour le paramètre par défaut de 0,95 de Khiops, MDLPC obtient une robustesse équivalente, mais est significativement dominée par Khiops en taux de bonne prédiction, alors que ChiSplit(0,95) à un taux de bonne prédiction équivalent, mais une robustesse significativement moins bonne. On remarque que la nouvelle version de Khiops améliore les performances de KhiopsBasic sur l'ensemble des deux critères, et que la post-optimisation des intervalles en aval de KhiopsRobust apporte une amélioration légère et systématique en taux de bonne prédiction. Khiops est très stable vis à vis de son paramétrage. Ses performances entre les paramètres 0,5 et 0,95 sont très stables. Un niveau de stabilité similaire est atteint dans ChiSplit entre les paramètres 0,99 et 0,999, et pour ChiMerge entre les paramètres 0,999 et 0,9999. Afin d'étudier précisément l'influence du paramétrage de la méthode Khiops, nous en avons évalué les performances pour des valeurs extrêmes et irréalistes de paramètre, de 0,99999 à 1e-9. Pour une valeur très proche de 1, la méthode est très « conservatrice » et ne produit que rarement des discrétisations multi-intervalles. De manière analogue à ChiMerge ou ChiSplit, cela entraîne une très bonne robustesse au détriment d'un faible taux de bonne prédiction. La performance prédictive s'améliore très rapidement quand la valeur du paramètre diminue, pour atteindre son maximum pour une plage de valeurs raisonnable, entre 0,95 et 0,5, avec très faible perte de robustesse. Au delà de ce seuil de 0,5 (qui correspond à environ 50% de chance de produire une discrétisation multi-intervalles pour un attribut indépendant de l'attribut cible), il y a sur-apprentissage, ce qui se traduit par une diminution lente de la robustesse. Par contre, on observe ici un comportement inhabituel : le sur-apprentissage n'entraîne pas de détérioration de la performance prédictive en test. Ce comportement étonnant a été observé même pour une valeur extrême de 1e-9 du paramétrage. Ce phénomène est très intéressant, car il sécurise l'utilisation du paramétrage de Khiops.

En conclusion, la méthode Khiops réalise les meilleurs compromis entre robustesse et taux de bonne prédiction quand on fait varier son paramétrage, et est très stable vis à vis de son paramétrage. Sur la plage de paramètres comprise entre 0,95 et 0,5, la méthode Khiops domine l'ensemble des autres méthodes testées indépendamment de leur paramétrage.

5.3.3 Performance prédictive et taille des discrétisations

La figure 9 présente l'ensemble des résultats sur le plan des critères de taux de bonne prédiction et de nombre d'intervalles.

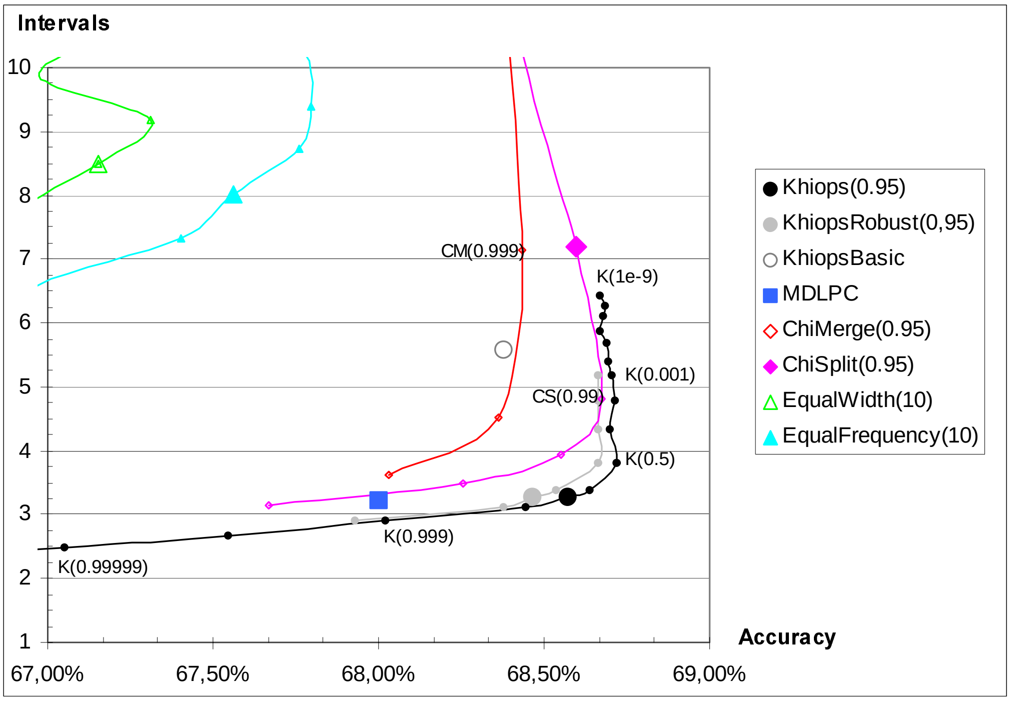


Figure 9 : Evaluation des méthodes de discrétisation pour les critères de taux de bonne prédiction et de taille des discrétisations

Les méthodes non supervisées EqualWidth et EqualFrequency sont très largement dominées par l'ensemble des autres méthodes sur les deux critères étudiés.

La méthode ChiMerge est la moins performante des méthodes supervisées testées, et obtient notamment des nombres d'intervalles très élevés. Ainsi, pour son paramétrage par défaut de 0,95, elle obtient en moyenne 66 intervalles.

La méthode MDLPC obtient de très bons résultats en termes de nombre d'intervalles, mais est dépassée en taux de bonne prédiction par les autres méthodes, notamment ChiSplit et Khiops sur des plages importantes de leur paramétrage. Ce comportement est conforme aux résultats de l'expérimentation sur la résistance au bruit, où l'on a vu que MDLPC avait un comportement très « conservateur », en ne prenant jamais le risque de faire une discrétisation abusive, même quand une légère information pourait émerger et se distinguer du bruit.

ChiSplit obtient de très bonnes performances, mais est très sensible à son paramétrage. Elle atteint des taux de bonne prédiction très élevés, mais nécessite systématiquement un nombre d'intervalles significativement plus élevé que Khiops pour des performances prédictives comparables.

Les deux courbes pour ChiMerge et ChiSplit ont des formes similaires comme pour l'analyse bi-critères précédente, la robustesse étant remplacée par le nombre d'intervalles. Quand on diminue la valeur de leur paramètre, le nombre d'intervalles augmente rapidement, et entraîne une détérioration du taux de bonne prédiction.

La méthode Khiops obtient les meilleurs résultats sur l'ensemble des deux critères, et sa courbe semble située sur la courbe optimale de Pareto pour l'ensemble des méthodes testées. Par exemple, pour le paramètre par défaut de 0,95 de Khiops, MDLPC obtient un nombre d'intervalles équivalent, mais est significativement surpassé par Khiops en taux de bonne prédiction, alors que ChiSplit(0,95) a un taux de bonne prédiction équivalent, mais un nombre d'intervalles plus de deux fois plus important. On remarque que la nouvelle version de Khiops améliore les performances de KhiopsBasic sur l'ensemble des deux critères, et que la post-optimisation des intervalles en aval de KhiopsRobust apporte une amélioration légère et systématique en taux de bonne prédiction. Khiops est très stable vis à vis de son paramétrage. Pour une valeur très proche de 1, la méthode est très « conservatrice » et ne produit que rarement des discrétisations multi-intervalles. La performance prédictive croît avec le nombre d'intervalles quand la valeur du paramètre diminue, pour atteindre son maximum pour une plage de valeurs comprise entre 0,95 et 0,5. Au delà de ce seuil de 0,5, il y a sur-apprentissage, ce qui se traduit par une augmentation lente du nombre d'intervalles. Le sur-apprentissage n'entraîne pas de détérioration de la performance prédictive en test. Ce comportement étonnant peut s'expliquer en se rapportant à l'étude sur la résistance au bruit. Le sur-apprentissage entraîné par une valeur « laxiste » du paramétrage (proche de 0) entraîne la production de nouveaux intervalles. Ces nouveaux intervalles sont inutiles (voire dommageables), mais ils sont produits dans des zones où l'algorithme identifie une « structure » émergeant légèrement du bruit. Ces structures correspondent à des « fluctuations statistiques » peu différenciées de leur voisinage, ayant un impact faible sur le taux de bonne prédiction. De plus, le cumul de ces fluctuations a tendance à se compenser, ce qui finalement n'entraîne pas de détérioration sur la performance prédictive de l'algorithme.

En conclusion, la méthode Khiops réalise les meilleurs compromis entre nombres d'intervalles et taux de bonne prédiction quand on fait varier son paramétrage, et est très stable vis à vis de son paramétrage. Sur la plage de paramètres comprise entre 0,95 et 0,5, la méthode Khiops domine l'ensemble des autres méthodes testées indépendamment de leur paramétrage.

5.3.4 Performance prédictive et robustesse du prédicteur Bayesien Naïf

L'apport de l'analyse multi-critères effectuée pour les prédicteurs univariés suggère de procéder à une analyse similaire pour les performances du prédicteur Bayesien Naïf. Nous avons donc procédé à des expérimentations similaires en mesurant le taux de bonne prédiction et la robustesse des méthodes sur les quinze jeux de données et en étudiant l'influence du paramétrage des méthodes. Malheureusement, le faible nombre de jeux de données rend cette fois les différences observées faiblement significatives. Ainsi, par exemple, l'analyse des différences significatives par le test de Student montre qu'il faut cette fois environ 2% de différences entre les performances prédictives de deux méthodes pour que cette différence soit significative à 95%, ce qui rend la plupart des méthodes apparemment équivalentes. La figure 10, dans laquelle on a remplacé les courbes des figures multi-critères précédentes par des nuages de points, est donc à interpréter avec prudence. On peut néanmoins observer certaines tendances générales, qui corroborent en partie les conclusions précédentes. La méthode EqualWidth est la moins performante des méthodes. Toutes les méthodes supervisées obtiennent des résultats de qualité prédictive équivalentes (y compris ChiMerge pour des valeurs de paramètre supérieures à 0,999). Les méthodes ChiMerge et ChiSplit sont très instables vis à vis de leur paramétrage. Khiops est stable pour un paramétrage variant de 0,5 à 0,95, et pour ces valeurs, elle obtient des résultats non inférieurs à ceux des autres méthodes supervisées. La méthode EqualFrequency paraît clairement dominer les autres méthodes, mais est relativement sensible à son paramétrage. Il parait surprenant que cette méthode non supervisée surpasse toutes les autres en taux de bonne prédiction et en robustesse. En fait, cette méthode est intrinsèquement très robuste, en produisant des intervalles ayant une fréquence minimale importante et équilibrée. De plus, elle respecte le prérequis du prédicteur Bayesien Naïf, qui se base sur l'hypothèse d'indépendance des attributs connaissant l'attribut cible, contrairement aux méthodes supervisées qui « cassent » cette hypothèse en liant chaque attribut descriptif à l'attribut cible, et donc en les corrélant entre eux.

En conclusion, la méthode Khiops n'est surpassée que par la méthode EqualFrequency, qui paraît la méthode la plus adaptée dans le cas du prédicteur Bayesien Naïf. On peut recommander de déterminer le nombre d'intervalles de EqualFrequency par validation croisée pour optimiser le paramétrage de cette méthode et ainsi tirer le meilleur parti du prédicteur Bayesien Naïf.

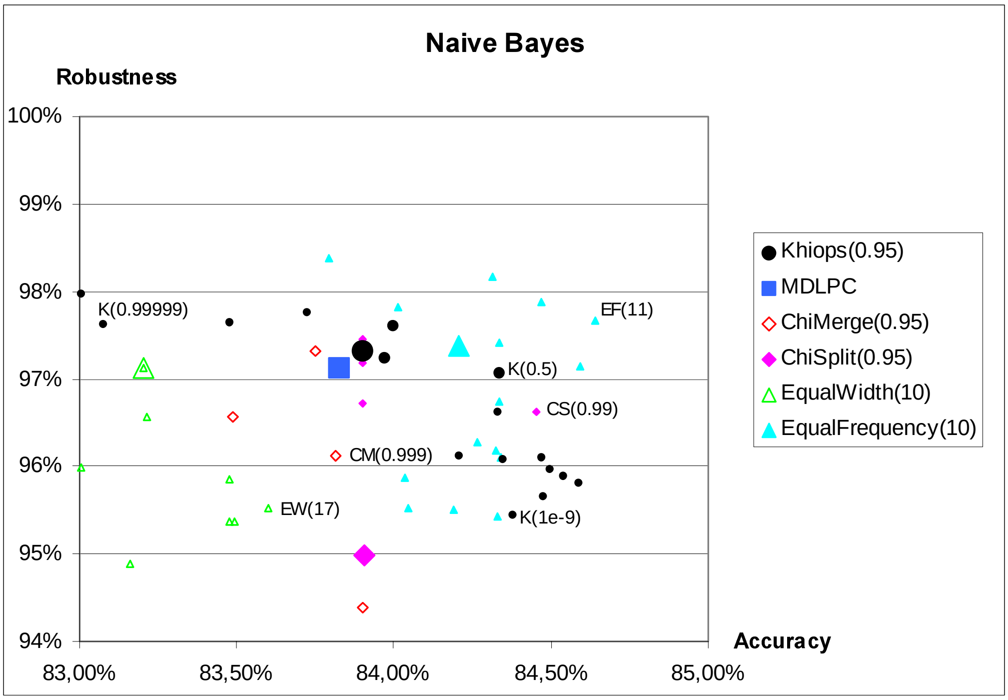


Figure 10 : Evaluation des méthodes de discrétisation en pré-processing pour un prédictuer Bayssien Naïf, pour les critères de taux de bonne prédiction et de robustesse

Conclusion

La méthode Khiops discrétise un attribut continu en minimisant la probabilité d'indépendance entre attribut discrétisé et attribut cible. Lors d'une discrétisation, de nombreuses fusions d'intervalles sont effectuées, donnant lieu à des variations DeltaKhi2 de la valeur du Khi2 du tableau de contingence. Nous avons montré que dans le cas d'un attribut descriptif indépendant d'un attribut cible à J modalités, ces variations du Khi2 suivent approximativement une loi du Khi2 à J-1 degrés de liberté. Cela nous a permis de modéliser la loi MaxDeltaKhi2 du plus grand DeltaKhi2 intervenant lors de la discrétisation d'un attribut indépendant de l'attribut cible. En imposant à l'algorithme d'accepter toute fusion entraînant une variation du Khi2 inférieure à ce MaxDeltaKhi2, la nouvelle version robuste de l'algorithme Khiops apporte alors la garantie que les attributs sans intérêt prédictif sont discrétisés en un seul intervalle 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 de données. 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.

Des expérimentations comparatives intensives ont été menées sur de nombreux jeux de données de l'UCI. Les résultats ont montré que la méthode Khiops était la plus performante sur tous les critères étudiés, à savoir le taux de bonne prédiction, la robustesse, le faible nombre d'intervalles produits et la résistance au bruit. Une analyse multi-critères des résultats s'est avérée très informative, en permettant d'une part d'évaluer la méthode Khiops de façon précise, et d'autre part d'améliorer la compréhension du problème de la discrétisation. Cette analyse a également révélé un comportement intéressant de la méthode Khiops, dont la performance prédictive ne décroît pas même quand son paramétrage la force à sur-apprendre.

En résumé, la méthode Khiops est actuellement la seule méthode de discrétisation basée sur un algorithme permettant de contrôler a priori le sur-apprentissage. Elle obtient des performances de premier plan tant en performance prédictive qu'en concision des discrétisations produites.

Références

Abramowitz, M. and Stegun, I. (1970), *Handbook of Mathematical Functions*, Dover Publications, Ninth Printing.

Blake, C.L. and 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.

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

Dougherty, J., Kohavi, R and Sahami, M. (1995). Supervised and Unsupervised Discretization of Continuous Features. *Proceedings of the Twelf International Conference on Machine Learning*, Los Altos, CA : Morgan Kaufmann, 194-202.

Fayyad, U. & Irani, K. (1992). On the handling of continuous-valued attributes in decision tree generation. *Machine Learning, 8*: 87-102.

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*, 123-128.

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

Liu, H., Hussain, F., Tan, C.L. and Dash, M. (2002). Discretization: An Enabling Technique. *Data Mining and Knowledge Discovery 6(4)*: 393-423.

Lechevallier, Y (1990). Recherche d’une partition optimale sous contrainte d’ordre total. Technical Report, INRIA.

Numerical Recipes in C : The Art of Scientific Programing, Copyright © 1988-1992 by Cambridge University Press.

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

Zighed, D.A., Rabaseda, S. & Rakotomalala, R. (1998). Fusinter: a method for discretization of continuous attributes for supervised learning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 6(33)*: 307-326.

Zighed, D.A. & Rakotomalala, R. (2000), *Graphes d’induction*. HERMES Science Publications, 327-359.

6 Annexe : Approximation du DeltaKhi2 pour la méthode Khiops

6.1 *Introduction*

La loi du Khi2 donne la probabilité que deux lois soient indépendantes pour une valeur du Khi2 et un nombre de degrés de liberté donné. L'algorithme Khiops part d'un tableau du Khi2 initial et fusionne les lignes de ce tableau tant que la probabilité d'indépendance diminue.

On s'intéresse ici aux probabilités d'indépendance faibles, pour des valeurs de Khi2 et des nombres de degrés de liberté très importants. Typiquement, l'application de l'algorithme Khiops sur des bases réelles d'environ un million d'individus dans le cadre d'études Data Mining entraîne des valeurs de Khi2 et des nombres de degrés de liberté de l'ordre de un million dans les étapes initiales de l'algorithme. Dans les phases finales, si les variables sources et cibles sont fortement corrélées, on obtient de très grandes valeurs du Khi2 pour de petits nombres de libertés, ce qui correspond à des probabilités d'indépendance pratiquement nulle (de l'ordre de 10^{10000}). L'algorithme Khiops basant son critère d'arrêt automatique sur l'évolution de la probabilité d'indépendance, il est nécessaire de pouvoir évaluer cette probabilité de façon suffisamment précise.

L'évaluation de la probabilité du Khi2 pour ces domaines de valeurs pose de nombreux problèmes numériques avec les méthodes habituelles Pour fixer les idées, sur une machine 32 bits, les réels sont stockés avec une précision de 15 chiffres pour la mantisse, et pour des exposants variant entre 10^{-308} et 10^{308}.

On ne sait pas calculer la loi du Khi2 pour des grandes valeurs du nombre de degrés de liberté. Les bibliothèques mathématiques évaluent la loi du Khi2 par un développement en série ou en fraction continue de la loi Gamma incomplète pour des petits nombres de degrés libertés (inférieur à quelques dizaines), et utilisent l'approximation gaussienne pour les grands nombres de degrés de liberté. Ces routines ne sont pas utilisables pour l'algorithme Khiops pour les bases de données réelles à cause des limites de l'exposant (vite atteintes), de la qualité relative de l'approximation gaussienne de la loi Gamma incomplète, et du problème de transition entre évaluation par la loi Gamma incomplète et par son approximation gaussienne.

Après un rappel de la définition de la loi du Khi2 dans la partie 1, on montre dans la partie 2 que la probabilité d'indépendance calculée avec la loi du Khi2 devient plus petite que ½ à partir des valeurs du Khi2 supérieure au nombre de degrés de liberté. Par la suite, on ne s'intéresse qu'à ce cas, le seul pertinent pour la méthode Khiops (qui peut continuer ses fusions de lignes de Khi2 tant que ce seuil n'est pas atteint). Dans la partie 3, on montre que l'on peut approximer le logarithme de la loi de probabilité du Khi2 sans problème numérique sur de grandes plages de valeurs. Le DeltaKhi2 est la diminution maximale de la valeur du Khi2 qui permet de diminuer la probabilité d'indépendance pour une diminution donnée du nombre de degrés de liberté. On montre dans la partie 4 que l'on peut approximer la valeur du DeltaKhi2 sans problème numérique. Enfin, dans la partie 5, on procède à quelques simulations numériques pour illustrer l'utilisation des approximations retenues.

6.2 *Loi du Khi2 et loi Gamma*

On rappelle la définition de la loi Gamma et de la loi du Khi2, et quelles unes de leurs propriétés qui seront utilisées dans ce document.

La fonction Γ est définie pour x > 0 par $\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}\,dt$

$\Gamma(x+1)=x\Gamma(x)$, en particulier, pour tout entier n, $\Gamma(n+1)=n!$

La loi Gamma d'indice α et ν est la loi définie sur $\Re^+$ par la densité $\gamma_{\alpha,\nu}(t)=\frac{1}{\Gamma(\nu)}\alpha^\nu t^{\nu-1}e^{-\alpha t}$

La loi $\gamma_{\frac{1}{2},\frac{n}{2}}$ est la loi du Khi2 a n degrés de liberté ou loi de Pearson. Sa densité est $\gamma_{\frac{1}{2},\frac{n}{2}}(t)=\frac{1}{\Gamma(n/2)}(1/2)^{\frac{n}{2}}t^{\frac{n}{2}-1}e^{-\frac{t}{2}}$.

Soit $Q(x,n)$ la probabilité que la valeur du Khi2 soit supérieure à x pour une loi du Khi2 à n degrés de liberté. $Q(x,n)=\int_x^\infty \gamma_{\frac{1}{2},\frac{n}{2}}(t)dt$

Par la suite, toutes les formules utilisées seront référencées par [AS§*number*]. Elles sont extraites du livre suivant :
« Handbook of Mathematical Functions », Milton Abramowitz and Irene Stegun, Dover Publications, Ninth Printing, 1970.

On a la formule de récurrence suivante :

$$Q(x,n+2)=Q(x,n)+\frac{(x/2)^{\frac{n}{2}}e^{-\frac{x}{2}}}{\Gamma(n/2+1)}\quad \text{[AS§26.4.8]}$$

En particulier, on a :

$$Q(x,2n)=e^{-\frac{x}{2}}\left(1+(x/2)+\frac{1}{2!}(x/2)^2+\ldots+\frac{1}{(n-1)!}(x/2)^{n-1}\right)\text{[AS§6.5.13]}$$

6.3 Equiprobabilité pour x=n

Proposition 1 : Q(n,n) converge vers ½ quand n tend vers l'infini. La différence entre ½ et Q(n,n) est de l'ordre de $\frac{1}{3\sqrt{\pi}\sqrt{n}}$.

Preuve :
La formule de Stirling permet d'approximer la loi Gamma.

$$\Gamma(x)\underset{x\to\infty}{\sim} e^{-x}x^{x-\frac{1}{2}}\sqrt{2\pi}\left(1+\frac{1}{12x}+\frac{1}{288x^2}-\frac{139}{51840x^3}-\frac{571}{2488320x^4}+\ldots\right)\quad\text{[AS§6.1.37]}$$

Pour la loi Gamma incomplète $\Gamma(a,x)=\int_x^{\infty}t^{a-1}e^{-t}dt$, on a :

$$\Gamma(x+1,x)\underset{x\to\infty}{\sim} e^{-x}x^{x}\left(\sqrt{\frac{\pi}{2}}x^{\frac{1}{2}}+\frac{2}{3}+\frac{\sqrt{2\pi}}{24}x^{-\frac{1}{2}}+\ldots\right)\text{[AS§6.5.35]}$$

Or $Q(x,n)=\frac{1}{\Gamma(n/2)}(1/2)^{\frac{n}{2}}\int_x^{\infty}t^{\frac{n}{2}-1}e^{-\frac{t}{2}}dt$

soit $Q(x,n)=\frac{1}{\Gamma(n/2)}\int_{x/2}^{\infty}t^{\frac{n}{2}-1}e^{-t}dt=\frac{\Gamma(n/2,x/2)}{\Gamma(n/2)}$

Ainsi, pour le couple de valeurs (n, n+2)

$$Q(n,n+2)=\frac{\Gamma(n/2+1,n/2)}{\Gamma(n/2+1)}=\frac{\Gamma(n/2+1,n/2)}{(n/2)\Gamma(n/2)}$$

En utilisant [AS§6.5.35] pour le numérateur et [AS§6.1.37] pour le dénominateur, on obtient :

$$Q(n,n+2)\underset{n\to\infty}{\sim}\frac{e^{-\frac{n}{2}}(n/2)^{\frac{n}{2}}\left(\sqrt{\frac{\pi}{2}}(n/2)^{\frac{1}{2}}+\frac{2}{3}+\frac{\sqrt{2\pi}}{24}(n/2)^{-\frac{1}{2}}+\ldots\right)}{(n/2)e^{-\frac{n}{2}}(n/2)^{\frac{n}{2}-\frac{1}{2}}\sqrt{2\pi}\left(1+\frac{1}{12}(n/2)^{-1}+\frac{1}{288}(n/2)^{-2}-\frac{139}{51840}(n/2)^{-3}-\frac{571}{2488320}(n/2)^{-4}+\ldots\right)}$$

$$Q(n,n+2)\underset{n\to\infty}{\sim}\frac{1}{2}\frac{\left(1+\frac{2\sqrt{2}}{3\sqrt{\pi}}(n/2)^{-\frac{1}{2}}+\frac{1}{12}(n/2)^{-1}+\ldots\right)}{\left(1+\frac{1}{12}(n/2)^{-1}+\frac{1}{288}(n/2)^{-2}-\frac{139}{51840}(n/2)^{-3}-\frac{571}{2488320}(n/2)^{-4}+\ldots\right)}$$

Par ailleurs, d'après la formule de récurrence [AS§26.4.8] de calcul de la loi du Khi2, on a

$$Q(n,n+2)=Q(n,n)+\frac{(n/2)^{\frac{n}{2}}e^{-\frac{n}{2}}}{\Gamma(n/2+1)}$$

c'est à dire $Q(n,n)=Q(n,n+2)-\frac{1}{2}\frac{\left(\sqrt{2\pi}(n/2)^{\frac{n}{2}-\frac{1}{2}}e^{-\frac{n}{2}}\right)\frac{\sqrt{2}}{\sqrt{\pi}}(n/2)^{-\frac{1}{2}}}{\Gamma(n/2)}$

En intégrant cet ajustement dans la formule générale, on a une diminution en valeur absolue du coefficient de $(n/2)^{-\frac{1}{2}}$ dans le numérateur de l'expression :

D'où $$\boxed{Q(n,n) \underset{n\to\infty}{\sim} \frac{1}{2}\,\frac{\left(1-\frac{\sqrt{2}}{3\sqrt{\pi}}(n/2)^{-\frac{1}{2}}+\frac{1}{12}(n/2)^{-1}+\dots\right)}{\left(1+\frac{1}{12}(n/2)^{-1}+\frac{1}{288}(n/2)^{-2}-\frac{139}{51840}(n/2)^{-3}-\frac{571}{2488320}(n/2)^{-4}+\dots\right)}}$$

Donc $Q(n,n)$ converge vers ½ quand n tend vers l'infini, et $\frac{1}{3\sqrt{\pi}\sqrt{n}}$ est le premier terme du développement de Q(n ,n).

Le résultat précédent montre le comportement asymptotique de Q(n,n). En pratique, Q(n,n) converge très rapidement vers ½. Ainsi, $Q(1,1)\approx 0{,}32,\ Q(10,10)\approx 0{,}44,\ Q(100,100)\approx 0{,}48$. On peut alors considérer que pour toute valeur de n, la probabilité d'indépendance liée à la loi du Khi2 est de l'ordre de ½ quand la valeur du Khi2 est égal au nombre de degrés de liberté.

6.4 Calcul du logarithme de probabilité du Khi2

Par la suite, on ne s'intéressera qu'au cas où x>n.

6.4.1 Calcul de ln(Q(x,1))

On passe par le complémentaire de la fonction d'erreur $erfc(x)=\frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}\,dt$ [AS§7.1.1]

On a $Q(x,1)=erfc\left(\sqrt{x/2}\right)$

Selon « Numerical Recipes in C », la fonction erfc(x) peut être évaluée par l'approximation de Chebyshev de la façon suivante : $erfc(x)\approx te^{\left(-x^2+\sum_{i=1}^{9}a_it^i\right)}$ avec $t=\frac{1}{1+x/2}$. L'erreur fractionnaire est de l'ordre de 10^{-7}.

On a ainsi une très bonne approximation de ln(Q(x,1)), y compris pour les grandes valeurs de x.

6.4.2 Calcul de ln(Q(x,2))

$$Q(x,2)=e^{-\frac{x}{2}}$$

$$\ln(Q(x,2))=-\frac{x}{2}$$

6.4.3 Calcul de ln(Q(x,n)) pour n > 2

On se base sur l'expression de Q(x,n) sous forme de fraction continue.

$$Q(x,n)=\frac{x^{\frac{n}{2}}e^{-\frac{x}{2}}}{2^{\frac{n}{2}}\Gamma(n/2)}FC(x,n) \text{ avec } FC(x,n)=\frac{1}{x/2+}\,\frac{1-n/2}{1+}\,\frac{1}{x/2+}\,\frac{2-n/2}{1+}\,\frac{1}{x/2+\dots} \qquad \text{[AS§26.4.10]}$$

Selon « Numerical Recipes in C », la fraction continue converge très rapidement pour x >n+2. Dans ce cas, la convergence demande de l'ordre de quelques fois $\sqrt{n/2}$ étapes, essentiellement quand x est proche de n.

En se basant sur la formule de récurrence [AS§26.4.8] de calcul de la loi du Khi2 $Q(x,n+2)=Q(x,n)+\frac{2}{n}\frac{x^{\frac{n}{2}}e^{-\frac{x}{2}}}{2^{\frac{n}{2}}\Gamma(n/2)}$, on peut se ramener au cas où la convergence est rapide dès que n > 2 et x > n.

En passant au logarithme, on a

$$\ln(Q(x,n)) = (n/2)\ln(x/2) - x/2 - \ln(\Gamma(n/2)) + \ln(FC(x,n))$$

Le logarithme de la fonction Gamma est approximé très précisément par

$$\ln(\Gamma(x)) \underset{x\to\infty}{\sim} (x - 1/2)\ln(x) - x + \ln(2\pi)/2 + \frac{1}{12x} - \frac{1}{360x^3} + \frac{1}{1260x^5} - \frac{1}{1680x^7} + \ldots \qquad \text{[AS§6.1.41]}$$

Le logarithme de Q(x,n) pourra ainsi être évalué avec précision.

Néanmoins, pour n très grand et x proche de n, le nombre d’étapes nécessaires au calcul de ln(Q(x,n)) devient important.

6.5 Calcul du DeltaKhi2

6.5.1 Introduction

Pour une valeur x de Khi2 et un nombre n de degrés de liberté, la loi du Khi2 permet d’évaluer la probabilité $Q(x,n)$ d’indépendance de deux variables pour une valeur de Khi2 supérieure à x. Si on diminue le nombre de degrés de liberté de k, on cherche à évaluer la nouvelle valeur x-dx de Khi2 permettant d’obtenir la même probabilité d’indépendance.

$$DK(x,n,k) = dx \Leftrightarrow Q(x,n) = Q(x - dx, n - k)$$

On vérifie aisément que $DK(x,n,k_1 + k_2) = DK(x,n,k_1) + DK(x - DK(x,n,k_1), n - k_1, k_2)$

Cette formule permet de calculer la valeur de DeltaKhi2 pour tout k dès que l’on sait la calculer pour k=1.

6.5.1.1 Cas où x est plus petit que n

Quand x est égal à n, on a vu que Q(x,n) est très proche de l’équiprobabilité. Quand x est plus petit que n, on ne cherchera pas à calculer la valeur de DeltaKhi2 (qui pose des problèmes numériques dans ce domaine de valeurs). On considérera en effet qu’une probabilité d’indépendance supérieure ou de l’ordre de 0,5 n’est pas intéressante en soit et que la méthode Khiops doit dans ce cas systématiquement accepter les fusions.

6.5.1.2 Cas où n est petit et x/n petit

Dans le cas où n est petit (typiquement inférieur à 30), on dispose d’approximations de ln(Q(x,n)) de très bonne qualité calculables rapidement et sans problème numérique (le passage au logarithme résout le problème des limites de l’exposant). Pour des valeurs de x/n petites (typiquement inférieures à 100), la valeur de ln(Q(x,n)) reste petite (de l’ordre de quelques centaines), ce qui permet d’obtenir une très bonne précision numérique avec la mantisse.

Il suffit alors de résoudre l’équation suivante d’inconnue dx.

$$DK(x,n,1) = dx \Leftrightarrow \ln(Q(x,n)) = \ln(Q(x - dx, n - 1)).$$

La fonction ln(Q(x,n)) étant monotone et la racine de l’équation étant comprise entre 0 et x, il est aisé de trouver la racine de l’équation par une méthode numérique d’approximation (par exemple recherche par dichotomie).

Cette méthode de calcul de DeltaKhi2 sera utilisée comme base de comparaison pour les méthodes d’approximation du DeltaKhi2 plus performantes. On pourra ainsi évaluer son domaine de validité.

6.5.1.3 Cas où n est grand ou x/n grand

Dans le cas où n est grand (supérieur à 30), voire très grand (typiquement de l’ordre de 1000000), l’approximation de ln(Q(x,n)) commence à poser des problèmes numériques, dus au temps de calcul de la fraction continue et à l’accumulation des erreurs numériques quand le nombre de termes évalués est trop grand, et à la précision limitée de la mantisse. Dans ce cas où l’on compare le comportement de ln(Q(x,n)) et ln(Q(x,n-1)), ces problèmes numériques dégradent rapidement la qualité des calculs de DeltaKhi2.

On utilise alors une nouvelle méthode de calcul de DeltaKhi2 pour k=2. On propose plusieurs bornes inf et sup de DeltaKhi2 pour k=2, ainsi qu’une méthode d’approximation du DeltaKhi2, qui ne pose pas de problèmes numériques y compris pour de très grandes valeurs de n ou de x.

Pour passer au calcul de DeltaKhi2 pour k=1, cette méthode est généralisée grâce à une nouvelle approximation valide jusqu’à environ n=1000. Pour n grand, on va montrer ci-dessous que $DK(x,n,1) \sim 1/2\, DK(x,n,2)$. On montrera par évaluation numérique la validité de cette approximation pour n supérieur à 1000 en comparant avec les valeurs de DeltaKhi2 obtenues avec la méthode de calcul pour n inférieur à 1000.

Montrons rapidement que $DK(x,n,1) \sim 1/2\, DK(x,n,2)$ quand n tend vers l’infini.

Posons, $f_{x,n}(y) = DK(x,n,ny)$ en passant au continu pour le troisième paramètre de DK .

Par définition, $DK(x,n,k) = dx \Leftrightarrow Q(x,n) = Q(x - dx, n - k)$

Il est facile de vérifier que $f_{x,n}(y)$ est une fonction différentiable, nulle en 0.

En prenant le développement de Taylor, $f_{x,n}(y) = yf'_{x,n}(0) + o(y^2)$.

On a alors $DK(x,n,1) = f_{x,n}(1/n) \sim 1/2\, f_{x,n}(2/n) = 1/2\, DK(x,n,2)$

6.5.2 Calcul de DeltaKhi2 pour un écart de 2 degrés de liberté

6.5.2.1 Bornes de DeltaKhi2

On cherche à évaluer $DK(x,n+2,2) = dx \Leftrightarrow Q(x,n+2) = Q(x-dx,n)$

Proposition 2 : $DK(x,n+2,2) = dx \Leftrightarrow Q(x-dx,n) - Q(x,n) = \dfrac{(x/2)^{\frac{n}{2}} e^{-\frac{x}{2}}}{\Gamma(n/2+1)}$

Ce résultat découle directement de la formule de récurrence de calcul de la loi Gamma incomplète [AS§26.4.8].

Corollaire : $DK(x,n+2,2) = dx \Leftrightarrow \displaystyle\int_{x-dx}^{x} t^{\frac{n}{2}-1} e^{-\frac{t}{2}} dt = \frac{2x^{\frac{n}{2}} e^{-\frac{x}{2}}}{n}$

Proposition 3 : $DK(x,n+2,2) \geq 2\ln(1+x/n)$

Preuve :

$$\int_{x-dx}^{x} t^{\frac{n}{2}-1} e^{-\frac{t}{2}} dt = \frac{2x^{\frac{n}{2}} e^{-\frac{x}{2}}}{n}$$

$$\frac{2x^{\frac{n}{2}} e^{-\frac{x}{2}}}{n} = \int_{x-dx}^{x} t^{\frac{n}{2}-1} e^{-\frac{t}{2}} dt \leq x^{\frac{n}{2}-1} \int_{x-dx}^{x} e^{-\frac{t}{2}} dt$$

$$\frac{2x^{\frac{n}{2}} e^{-\frac{x}{2}}}{n} \leq 2x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \left(e^{\frac{dx}{2}} - 1 \right)$$

$$x/n \leq \left(e^{\frac{dx}{2}} - 1 \right)$$

$$2\ln(1+x/n) \leq dx$$

Proposition 4 : $DK(x,n+2,2) \leq 2x/n$ pour x>n et n>2.

Preuve :

$$\int_{x-dx}^{x} t^{\frac{n}{2}-1} e^{-\frac{t}{2}} dt = \frac{2x^{\frac{n}{2}} e^{-\frac{x}{2}}}{n}$$

Soit $f(t) = t^{\frac{n}{2}-1} e^{-\frac{t}{2}}$. $f'(t) = \dfrac{1}{2} t^{\frac{n}{2}-2} e^{-\frac{t}{2}} (n-2-t)$

La fonction f est donc croissante avant n-2, puis décroissante.

Supposons que $x - dx \leq n-2$. Alors f est croissante de x-dx à n-2 puis décroissante de n-2 à x.

$$\int_{x-dx}^{x} t^{\frac{n}{2}-1} e^{-\frac{t}{2}} dt \geq \int_{n-2}^{x} t^{\frac{n}{2}-1} e^{-\frac{t}{2}} dt$$

$$\int_{x-dx}^{x} t^{\frac{n}{2}-1} e^{-\frac{t}{2}} dt \geq x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \left(x-(n-2)\right)$$

$$\frac{2x^{\frac{n}{2}} e^{-\frac{x}{2}}}{n} \geq x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \left(x-(n-2)\right)$$

$$\frac{2x}{n} \geq x-(n-2)$$

$$0 \geq (x-n)(n-2)$$

Ce dernier résultat étant faux dans le cas où x>n et n > 2, l'hypothèse $x-dx \leq n-2$ est par conséquent absurde.

Donc $x-dx > n-2$

f est donc décroissante sur l'intervalle d'intégration.

$$\int_{x-dx}^{x} t^{\frac{n}{2}-1} e^{-\frac{t}{2}} dt \geq x^{\frac{n}{2}-1} e^{-\frac{x}{2}} dx$$

$$\frac{2x^{\frac{n}{2}} e^{-\frac{x}{2}}}{n} \geq x^{\frac{n}{2}-1} e^{-\frac{x}{2}} dx$$

$$2x/n \geq dx$$

Proposition 5 : $DK(x, n+2, 2) \leq 2\ln(1+ex/n)$ pour x>n et n>2.

Preuve :

$t^{\frac{n}{2}-1}$ est une fonction croissante de t.

En se basant sur la borne précédente, on a $x-dx \geq x-2x/n$.

$$\frac{2x^{\frac{n}{2}} e^{-\frac{x}{2}}}{n} = \int_{x-dx}^{x} t^{\frac{n}{2}-1} e^{-\frac{t}{2}} dt \geq (x-2x/n)^{\frac{n}{2}-1} \int_{x-dx}^{x} e^{-\frac{t}{2}} dt$$

$$\frac{2x^{\frac{n}{2}} e^{-\frac{x}{2}}}{n} \geq 2x^{\frac{n}{2}-1} (1-2/n)^{\frac{n}{2}-1} e^{-\frac{x}{2}} \left(e^{\frac{dx}{2}}-1\right)$$

$$\frac{x}{n(1-2/n)^{\frac{n}{2}-1}} \geq e^{\frac{dx}{2}}-1$$

$$2\ln\left(1+\frac{x}{n(1-2/n)^{\frac{n}{2}-1}}\right) \geq dx$$

Soit $h(n)=\left(1-\frac{2}{n}\right)^{\frac{n}{2}-1}$. $h'(n)=\frac{n}{2}\left(1-\frac{2}{n}\right)^{\frac{n}{2}-1}\left(1+\ln\left(1-\frac{2}{n}\right)\right)$. Donc h est décroissante et positive. Sa limite est 1/e quand n tend vers l'infini.

Donc $2\ln(1+ex/n) \geq dx$

Corollaire 6 : $2\ln(1+x/n) \leq DK(x, n+2, 2) \leq 2\ln(1+ex/n) \leq 2\ln(1+x/n)+2$ pour x>n et n>2

Proposition 7 : $2\ln(1+x/n)\le DK(x,n+2,2)\le 2\ln(1+x/n(1+\varepsilon_n(x)))$ avec

$$\varepsilon_n(x)=(1-(2/x)\ln(1+e\,x/n))^{\frac{n}{2}-1}-1 \text{ pour x>n et n>2}$$

Preuve :

$t^{\frac{n}{2}-1}$ est une fonction croissante de t et $x-dx\ge 2\ln(1+e\,x/n)$.

On arrive au résultat proposé en utilisant le même principe de Preuve que pour la borne précédente.

Corollaire 8 : $DK(x,n+2,2)$ se comporte asymptotiquement comme $2\ln(1+x/n)$ quand x tend vers l'infini.

6.5.2.2 Amélioration des bornes pour x proche de n

Proposition 9 :Pour $n-2\le x\le n-2+\sqrt{2(n-2)}$

$DK(x,n+2,2)\ge(2x/n)\dfrac{2}{1+\sqrt{1+2x/n-2+4/n}}$ et pour $\left(n-2+\sqrt{2(n-2)}\right)(1+2/n)\le x$, cette borne inf devient une borne sup.

Preuve :

Soit $f(t)=t^{\frac{n}{2}-1}e^{-\frac{t}{2}}$

$$f'(t)=\frac{1}{2}t^{\frac{n}{2}-2}e^{-\frac{t}{2}}(n-2-t)$$

$$f''(t)=\frac{1}{4}t^{\frac{n}{2}-2}e^{-\frac{t}{2}}\left((t-(n-2))^2-2(n-2)\right)$$

La fonction f a donc son maximum en n-2, et possède un point d'inflexion de part et d'autre de son maximum.

	0		$n-2-$ $\sqrt{2(n-}$		$n-2$		$n-2+$ $\sqrt{2(n-}$		∞
$f''(t)$	0	+	0		-		0	+	0
$f'(t)$	0	↗	+	↘	0	↘	-	↗	0
$f(t)$	0		↗		+		↘		0

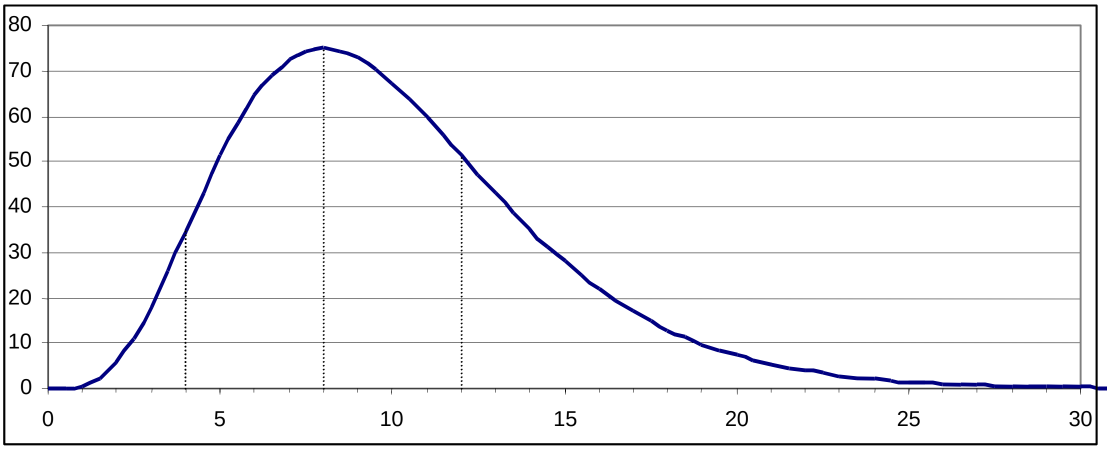


Figure 11 : Courbe f(t) pour n=10

Pour $n-2 \le x \le n-2+\sqrt{2(n-2)}$, la courbe entre x-dx et x se situe au-dessous de sa tangente en x. Pour $n-2+\sqrt{2(n-2)} \le x-dx$, elle se situe au-dessus de sa tangente en x. Compte tenu de $dx \le 2x/n$, on vérifie que ce dernier cas est vérifié dès que $\left(n-2+\sqrt{2(n-2)}\right)(1+2/n) \le x$.

L'intégrale T(x) entre x-dx et x de la tangente en x fournit donc respectivement un majorant puis un minorant de l'intégrale de $f(t)$ entre x-dx et x.

$$T(x) = \int_{x-dx}^{x} \left(f(x)+(t-x)f'(x)\right)dt$$

$$T(x) = dx.x^{\frac{n}{2}-1}e^{-\frac{x}{2}} + \frac{1}{4}dx^2 x^{\frac{n}{2}-2}e^{-\frac{x}{2}}\left(x-(n-2)\right)$$

Avant le point d'inflexion, on a

$$T(x) \ge \int_{x-dx}^{x} t^{\frac{n}{2}-1}e^{-\frac{t}{2}}dt = \frac{2x^{\frac{n}{2}}e^{-\frac{x}{2}}}{n}$$

$$\frac{1}{4}\left(x-(n-2)\right)dx^2 + x.dx - \frac{2x^2}{n} \ge 0$$

On a une équation du second degré d'inconnue dx.

Son déterminant est $\Delta = x^2\left(1+2\left(x-(n-2)\right)/n\right)$

Ses racines sont $dx = (2x/n)\dfrac{-1\pm\sqrt{1+2\left(x-(n-2)\right)/n}}{\left(x-(n-2)\right)/n}$

Pour x supérieur à n, l'équation admet donc une racine négative et une racine positive, et est négative en dx=0. Pour respecter l'inégalité, dx doit donc être supérieur à la racine positive de l'équation.

Donc $dx \ge (2x/n)\dfrac{-1+\sqrt{1+2\left(x-(n-2)\right)/n}}{\left(x-(n-2)\right)/n}$

$$dx \ge (2x/n)\frac{2}{1+\sqrt{1+2\left(x-(n-2)\right)/n}}$$

Légèrement au-dessus du point d'inflexion, pour $\left(n-2+\sqrt{2(n-2)}\right)(1+2/n) \le x$, cette borne inf devient une borne sup de dx.

Proposition 10 : $DK(n,n,2)$ converge vers 2 quand n tend vers l'infini

Preuve :

D'après la proposition 4, on a $DK(n,n,2) \leq \frac{2n}{n-2}$.

D'après la proposition 9, on a $\frac{2n}{n-2}\left(\frac{2}{1+\sqrt{1+8/(n-2)}}\right) \leq DK(n,n,2)$.

Cet encadrement assure la convergence de $DK(n,n,2)$ vers 2 quand n tend vers l'infini.

6.5.2.3 Approximation de DeltaKhi2

Soit $A(x,y,n) = \sum_{k=0}^{\infty} \frac{(n-2)(n-4)..(n-2k)}{x^k}\left(e^{\frac{y}{2}} e_k(-y/2) - 1\right)$

Avec $e_k(y) = \sum_{i=0}^{k} \frac{y^i}{i!}$ (développement limité à l'ordre k de e^y).

Proposition 11 : $DK(x,n+2,2) = dx \Leftrightarrow A(x,dx,n) = x/n$

Preuve :

On se base sur $DK(x,n+2,2) = dx \Leftrightarrow Q(x-dx,n) - Q(x,n) = \frac{(x/2)^{\frac{n}{2}} e^{-\frac{x}{2}}}{\Gamma(n/2+1)}$

En se ramenant à la définition de Q(x,n), on a :

$$Q(x-dx,n) - Q(x,n) = \frac{\Gamma(n/2, x/2 - dx/2) - \Gamma(n/2, x/2)}{\Gamma(n/2)}$$

D'après [AS§6.5.30], on a :

$$\Gamma(a,x+y) - \Gamma(a,x) = e^{-x} x^{a-1} \sum_{k=0}^{\infty} \frac{(a-1)(a-2)..(a-k)}{x^k}\left(e^{-y} e_k(y) - 1\right) \qquad (|y| < |x|)$$

Donc $Q(x-dx,n) - Q(x,n) = \frac{1}{\Gamma(n/2)} e^{-\frac{x}{2}} (x/2)^{\frac{n}{2}-1} \sum_{k=0}^{\infty} \frac{(n/2-1)(n/2-2)..(n/2-k)}{(x/2)^k}\left(e^{\frac{dx}{2}} e_k(-dx/2) - 1\right)$

c'est à dire $Q(x-dx,n) - Q(x,n) = \frac{1}{\Gamma(n/2)} e^{-\frac{x}{2}} (x/2)^{\frac{n}{2}-1} A(x,dx,n)$

Donc $\frac{(x/2)^{\frac{n}{2}} e^{-\frac{x}{2}}}{\Gamma(n/2+1)} = \frac{1}{\Gamma(n/2)} e^{-\frac{x}{2}} (x/2)^{\frac{n}{2}-1} A(x,dx,n)$

On cherche donc dx, solution de l'équation $A(x,dx,n) = x/n$

Evaluation numérique de $A(x,dx,n)$

$$A(x,y,n) = \sum_{k=0}^{\infty} \frac{(n-2)(n-4)..(n-2k)}{x^k}\left(e^{\frac{y}{2}} e_k(-y/2) - 1\right)$$

<u>Premier terme :</u> $\frac{(n-2)(n-4)..(n-2k)}{x^k}$

On est dans le cas où x > n. Le premier terme converge vers 0 plus vite que le terme d'une suite géométrique de raison k/n tant que k<n. Ce terme converge vers 0 d'autant plus vite que x/n est grand.

Pour n pair, ce terme devient nul à partir du rang k. Pour n impair, ce terme commence à augmenter en valeur absolue dès que k > n/2

Second terme : $\left(e^{\frac{y}{2}} e_k(-y/2) - 1 \right)$

La série $e_k(y) = \sum_{i=0}^{k} \frac{y^i}{i!}$ converge rapidement vers e^y dès que i > y, ce qui assure une convergence vers 0 des termes $e^{y/2} e_k(-y/2) - 1$. Il suffit de vérifier que l'évaluation de $y^i/i!$ ne pose pas de problème numérique. D'après le calcul des bornes de DeltaKhi2, on sait que $dx/2 \le \ln(1 + e.x/n)$. Ainsi, même dans un cas extrême comme x/n=1000000, on a dx/2<15. Or $15^{15}/15! \le 350000$. On reste donc très en deçà des problèmes numériques.

Convergence globale :

On s'intéresse au cas où n est impair et ou k > n/2 et k > y/2. Le premier terme se comporte comme $\frac{(n/2)!(k-n/2)!}{(x/2)^k}$. Le second terme, qui peut s'écrire sous la forme $e^{y/2}\left(e_k(-y/2) - e^{-y/2}\right)$, se comporte comme $e^{y/2}\frac{(y/2)^k}{k!}$. Le produit des deux termes se comporte approximativement comme $e^{y/2}\frac{(n/2)!(k-n/2)!}{k!}(y/x)^k$ Comme x>y, la série converge donc plus vite qu'une suite de raison y/x.

Dans les domaines de valeurs étudiés (n grand, x > n, $dx \le \ln(1 + e.x/n)$, la série $A(x, dx, n)$ converge très rapidement, avec un nombre d'itérations de l'ordre de quelques fois y.

Résolution de $A(x, dx, n) = x/n$

$A(x, dx, n) = \left(Q(x - dx, n) - Q(x, n)\right)\Gamma(n/2)e^{\frac{x}{2}}(x/2)^{1-\frac{n}{2}}$ est une fonction croissante de dx sur l'intervalle [0, x], valant 0 en 0. Il est alors facile de résoudre l'équation $A(x, dx, n) = x/n$ pour calculer dx, par exemple par recherche dichotomique. Les bornes calculées pour DeltaKhi2 permettent de restreindre l'intervalle de recherche, notamment la borne sup $dx/2 \le \ln(1 + e.x/n)$ qui permet de travailler un intervalle pour lequel la série $A(x, dx, n)$ converge rapidement.

6.5.3 Calcul de DeltaKhi2 pour un écart de 1 degré de liberté

On va utiliser exactement les mêmes principes que pour le calcul de DeltaKhi2 avec un écart de 2 degrés de liberté. Dans ce dernier cas, la formule de récurrence de calcul de Q(x,n) nous permet d'avoir la valeur exacte pour un saut de 2 degrés de liberté. On va ici remplacer cette valeur exacte par une approximation quand le saut est de un degré de liberté.

Proposition 12 : $Q(x, n+1) = Q(x, n) + \frac{(x/2)^{\frac{n}{2}} e^{-\frac{x}{2}}}{\Gamma(n/2)} DFC(x, n)$ avec

$$DFC(x, n) = \sqrt{x/2}\,\frac{\Gamma(n/2)}{\Gamma((n+1)/2)} FC(x, n+1) - FC(x, n)$$

Preuve :

Cela provient simplement de l'expression de Q(x,n) sous forme de fraction continue.

$Q(x, n) = \frac{x^{\frac{n}{2}} e^{-\frac{x}{2}}}{2^{\frac{n}{2}}\Gamma(n/2)} FC(x, n)$ avec $FC(x, n) = \frac{1}{x/2+}\,\frac{1-n/2}{1+}\,\frac{1}{x/2+}\,\frac{2-n/2}{1+}\,\frac{1}{x/2+\ldots}$

La fraction continue converge très rapidement pour x >n+2. Dans ce cas, la convergence demande de l'ordre de quelques fois $\sqrt{n/2}$ étapes, seulement dans le cas où x est proche de n.

Cela rend cette évaluation intéressante jusqu'à n de l'ordre de 1000.

Remarque : D'après la formule de Wallis [AS§6.1.49], $\frac{\Gamma(n/2)}{\Gamma((n+1)/2)} \underset{n\to\infty}{\sim} \frac{1}{\sqrt{n/2}}$.

Corollaire 13 : $DK(x,n+1,1)=dx \Leftrightarrow Q(x-dx,n)-Q(x,n)=\frac{(x/2)^{\frac{n}{2}}e^{-\frac{x}{2}}}{\Gamma(n/2)}DFC(x,n)$

Corollaire 14 : $DK(x,n+1,1)=dx \Leftrightarrow \int_{x-dx}^{x} t^{\frac{n}{2}-1}e^{-\frac{t}{2}}dt = x^{\frac{n}{2}}e^{-\frac{x}{2}}DFC(x,n)$

On rappelle que pour un écart de deux degrés de liberté, on a $DK(x,n+2,2)=dx \Leftrightarrow \int_{x-dx}^{x} t^{\frac{n}{2}-1}e^{-\frac{t}{2}}dt=\frac{2x^{\frac{n}{2}}e^{-\frac{x}{2}}}{n}$. Les deux formules sont similaires, et la plupart des bornes obtenues précédemment peuvent être transposées simplement au cas d'un écart de un degré de liberté.

On obtient alors les propositions suivantes.

Proposition 15 : $DK(x,n+1,1)\geq 2\ln(1+(x/2)DFC(x,n))$

Proposition 16 : $DK(x,n+1,1)\leq xDFC(x,n)$ pour x>n et n>2

Proposition 17 : $DK(x,n+1,1)\leq 2\ln\left(1+\frac{xDFC(x,n)}{2(1-DFC(x,n))^{\frac{n}{2}-1}}\right)$ pour x>n et n>2

Proposition 18 : $DK(x,n+1,1)=dx \Leftrightarrow A(x,dx,n)=(x/2)DFC(x,n)$ pour x>n et n>2

Preuve :

On cherche dx tel que

$$Q(x,n+1)-Q(x,n)=Q(x-dx,n)-Q(x,n)$$

Le résultat provient des deux formules suivantes.

$$Q(x,n+1)-Q(x,n)=\frac{(x/2)^{\frac{n}{2}}e^{-\frac{x}{2}}}{\Gamma(n/2)}DFC(x,n)$$

$$Q(x-dx,n)-Q(x,n)=\frac{1}{\Gamma(n/2)}e^{-\frac{x}{2}}(x/2)^{\frac{n}{2}-1}A(x,dx,n)$$

Résolution de $A(x,dx,n)=(x/2)DFC(x,n)$

Cette équation se résout de façon identique à l'équation $A(x,dx,n)=x/n$, pourvu que l'on se place dans le domaine de convergence rapide de DFC(x,n) (ce qui est le cas pour n <= 1000).

6.6 Evaluation numérique

6.6.1 Ln(Q(x,n)

Pour des raisons de lisibilité, on a utilisé l'opposé du logarithme base 10 de la probabilité du Khi2 :

ProbLevel(x,n) = -ln(Q(x,n)/ln(10)

On a tracé la courbe ProbLevel en fonction du nombre de degrés de liberté pour des valeurs de x proportionnelles à ce nombre de degrés de liberté. Pour couvrir un très large domaine de valeur, on a tracé les courbes pour ndl variant de 1 à 1000000000, et pour des ratios x/ndl variant de 1 à 1000000.

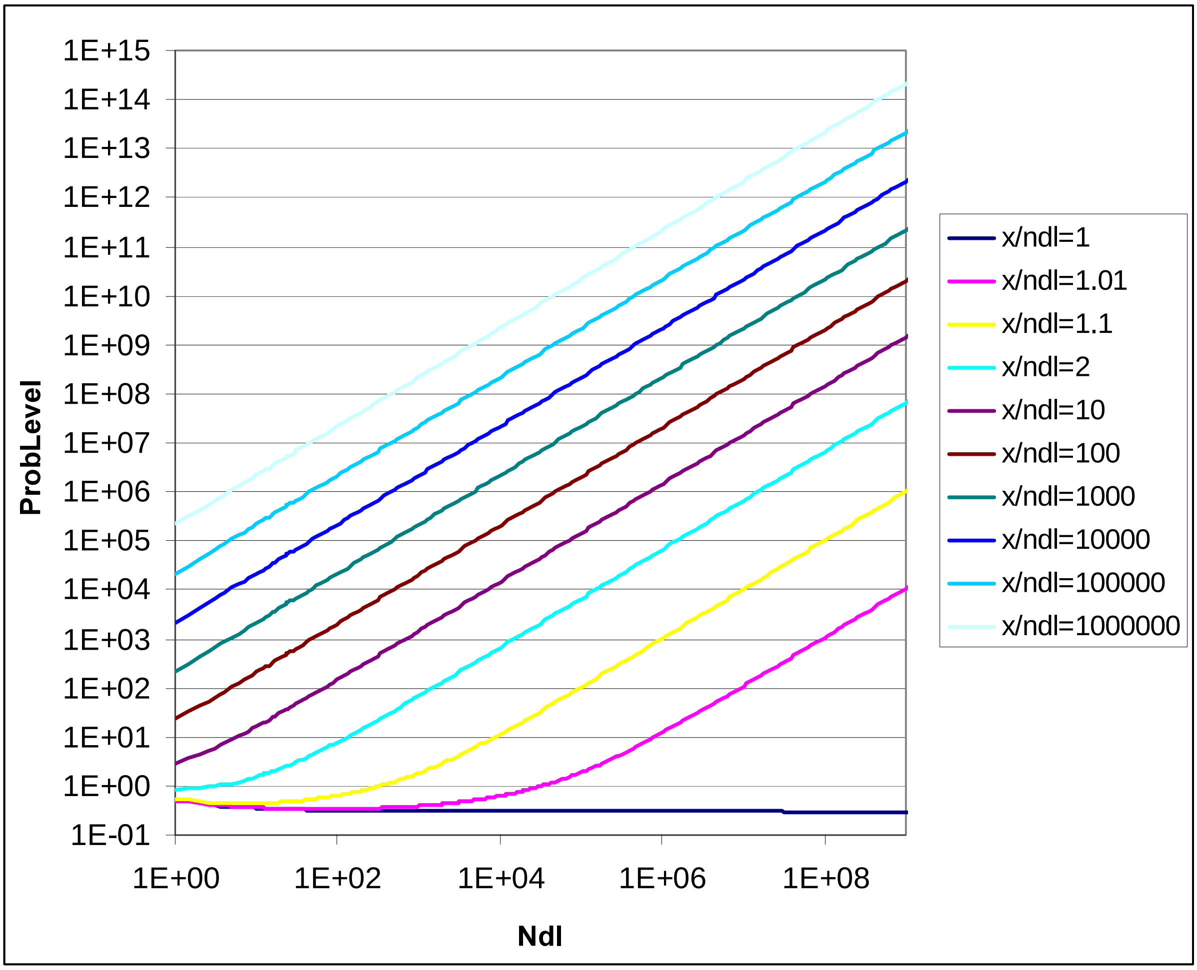


Figure 12 : Logarithme de la probabilité du Khi2 fonction du nombre de degrés de liberté, pour différents ratios x/ndl

On observe que conformément à ce qui a été démontré, la probabilité obtenue pour x/ndl=1 converge très rapidement vers 0,5. Quand x/ndl croit, il faut attendre une valeur de ndl assez importante pour se détacher de ce comportement d'équiprobabilité. Pour des ratios x/ndl plus importants, la probabilité d'indépendance diminue extrêmement vite avec le nombre de degrés de liberté. Les limites informatiques de l'exposant (10^{-308}) sont très rapidement dépassées, par exemple pour les valeurs (x/ndl=10 ;ndl=1000) ou (x/ndl=100 ;ndl=1000). On peut atteindre des niveaux de probabilité d'indépendance extrêmement bas (de l'ordre de $10^{-1000000}$ par exemple pour (x/ndl=100000 ;ndl=100)).

Le comportement des évaluations numériques est conforme à celui des méthodes habituelles pour les petites plages de valeurs, et paraît régulier par la suite.

6.6.2 Comparaison de plusieurs méthodes d'approximation de DeltaKhi2

On a comparé les trois méthodes d'approximation suivante pour le calcul du DeltaKhi2 :

- Khi2 Approx : Résolution de l'équation $\ln(Q(x-dx,n-1))=\ln(Q(x,n))$
- DK_1 Approx : Résolution de l'équation $A(x,dx,n)=(x/2)DFC(x,n)$
- DK_2/2 Approx : Résolution de l'équation $A(x,dx_2,n)=x/n$ et $dx=dx_2/2$

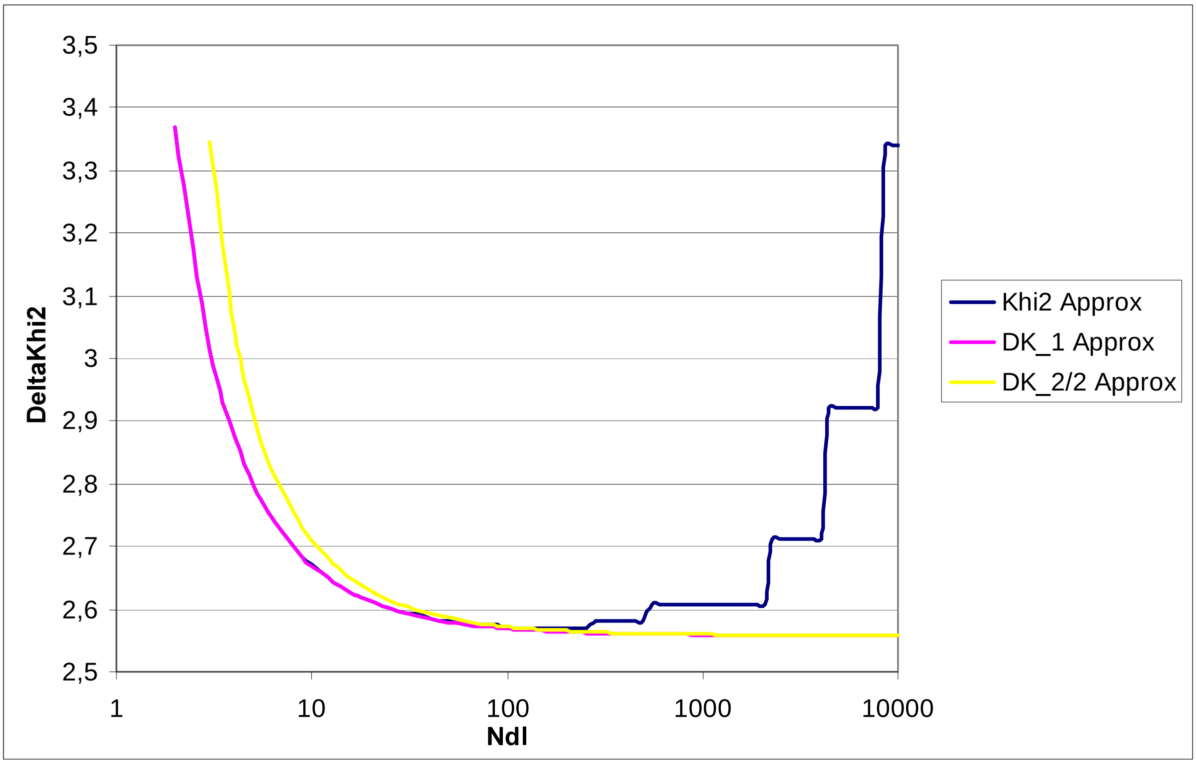


Figure 12 : Comparaison de trois méthodes d'approximation de DeltaKhi2 pour x/ndl=10

La première méthode sert de référence pour les petites valeurs de nombre de liberté. Au delà de ndl=100, son comportement numérique devient chaotique. La deuxième méthode coïncide avec la première pour les petites valeurs de ndl, puis suit un comportement conforme aux bornes calculées. Elle est numériquement très stable (même pour des valeurs de ndl de l'ordre de 1000000 non représentées ici), mais le nombre d'itérations nécessaire au calcul de $DFC(x,n)$ devient assez important pour des grands ndl. La troisième méthode n'est qu'une approximation grossière de DeltaKhi2 pour les petites valeurs de ndl. Elle permet néanmoins de confirmer la conjoncture $DK(x,n,1)\approx 1/2\,DK(x,n,2)$ pour les grandes valeurs de ndl. Cette conjoncture a été vérifiée numériquement pour de nombreux ratios x/ndl. Son utilisation n'est pas nécessaire, compte tenu de la fiabilité de la deuxième méthode. Elle permet cependant d'accélérer le temps de calcul de DeltaKhi2 pour les grandes valeurs de ndl.

6.6.3 DK(x,n,1)

On a tracé la courbe DeltaKhi2 en fonction du nombre de degrés de liberté pour des valeurs de x proportionnelles à ce nombre de degrés de liberté. Pour couvrir un très large domaine de valeur, on a tracé les courbes pour ndl variant de 1 à 1000000000, et pour des ratios x/ndl variant de 1 à 1000000.

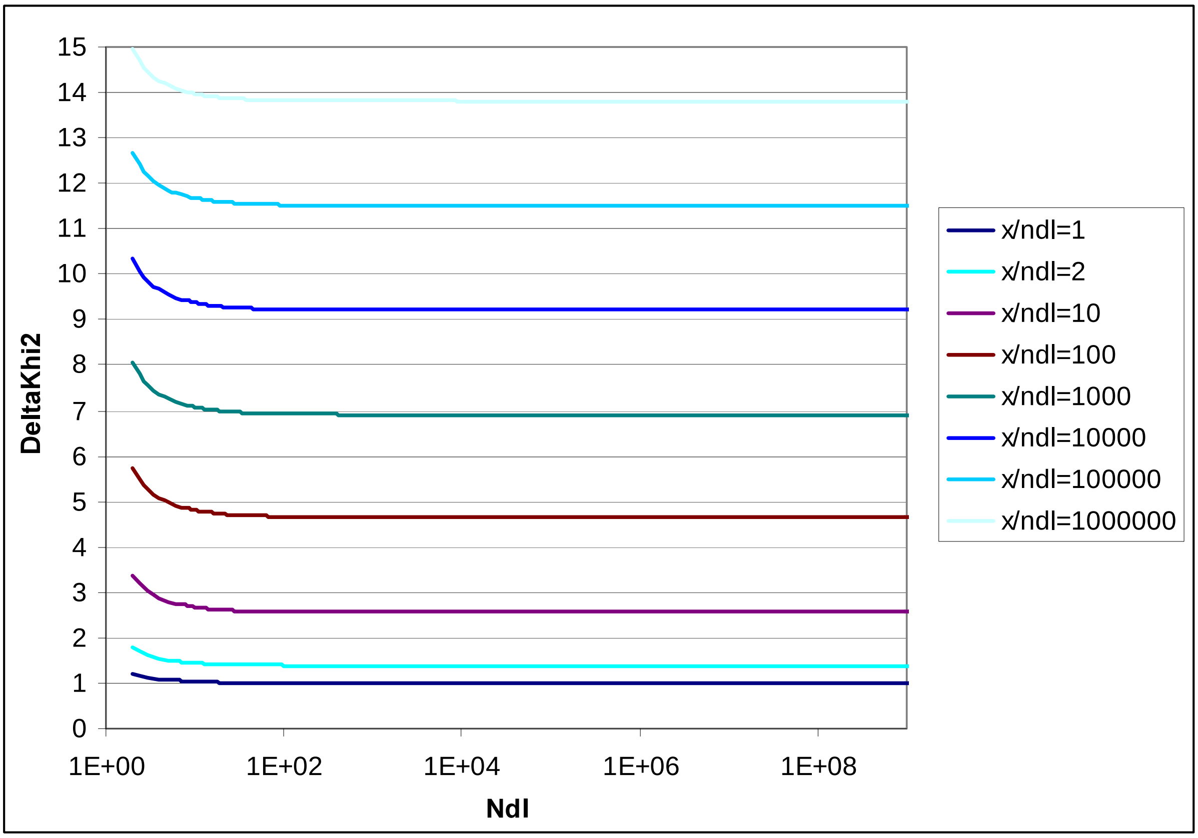


Figure 13 : DeltaKhi2 fonction du nombre de degrés de liberté, pour différents ratios x/ndl

Pour x/ndl = 1, la valeur de DeltaKhi2 est de l'ordre de 1. Cette valeur constitue donc la valeur minimum des valeurs de DeltaKhi2 utilisées par la méthode Khiops. Contrairement à la loi probabilité du Khi2 qui varie extrêmement vite en fonction de ses paramètres, la loi du DeltaKhi2 est très stable, et converge vite vers une valeur cible pour un ratio x/ndl donné. La valeur de DeltaKhi2 paraît stabilisée dès que le nombre de degrés de liberté dépasse quelques dizaines.

Les valeurs asymptotiques de DeltaKhi2 sont conformes au comportement asymptotique $2\ln(1+x/n)$ démontré pour les écarts de deux degrés de liberté. On observe ici une valeur plus petite de moitié, ce qui conforme la conjecture $DK(x,n,1)\approx 1/2\,DK(x,n,2)$ sur de grandes plages de valeur.

La valeur du DeltaKhi2 augmente très lentement avec la valeur du ratio x/ndl.

Le comportement des évaluations numériques de DeltaKhi2 ne présente aucune anomalie détectable sur de très grandes plages de valeur.

6.7 *Exemples de fusions*

La fusion de deux lignes entraîne un $DeltaKhi2 = -\frac{nn'}{n+n'}\sum_j \frac{(a_j - b_j)^2}{p_j}$

Prenons le cas de deux modalités cibles.

Soient $p_1 = p$ $p_2 = 1-p$
$a_1 = a$ $a_2 = 1-a$
$b_1 = b$ $b_2 = 1-b$

Dans ce cas, on a $\boxed{DeltaKhi2 = -\frac{nn'}{n+n'}\frac{(a-b)^2}{p(1-p)}}$

On choisit p=0,5 et on va calculer dans le tableau 14 les valeurs du DeltaKhi2 pour les fusions de deux lignes de Khi2 ayant toutes les combinaisons possibles d'effectifs observés de 0, 1, 2, 10, 11. En raison des symétries des combinaisons, seule une partie des colonnes du tableau est ici présente. Les cases en gris foncé représentent les fusions ayant un DeltaKhi2 nul. Les case en gris clair représentent les fusions ayant un DeltaKhi2 d'amplitude inférieure à 1.

Tableau 14 : Valeurs de DeltaKhi2 pour p=0,5 et pour des effectifs observés de 0, 1, 2, 10 et 11

	0-1	0-2	0-10	0-11	1-1	1-2	1-10	1-11	2-2	2-10	2-11	10-10	10-11	11-11
0-1	0,000	0,000	0,000	0,000	-0,667	-0,333	-0,030	-0,026	-0,800	-0,103	-0,088	-0,952	-0,866	-0,957
0-2	0,000	0,000	0,000	0,000	-1,000	-0,533	-0,056	-0,048	-1,333	-0,190	-0,164	-1,818	-1,656	-1,833
0-10	0,000	0,000	0,000	0,000	-1,667	-1,026	-0,173	-0,152	-2,857	-0,606	-0,535	-6,667	-6,144	-6,875
0-11	0,000	0,000	0,000	0,000	-1,692	-1,048	-0,182	-0,159	-2,933	-0,638	-0,564	-7,097	-6,548	-7,333
1-0	-2,000	-2,667	-3,636	-3,667	-0,667	-1,333	-3,030	-3,103	-0,800	-2,564	-2,659	-0,952	-1,048	-0,957
1-1	-0,667	-1,000	-1,667	-1,692	0,000	-0,133	-1,133	-1,190	0,000	-0,762	-0,831	0,000	-0,004	0,000
1-2	-0,333	-0,533	-1,026	-1,048	-0,133	0,000	-0,554	-0,600	-0,190	-0,267	-0,314	-0,290	-0,214	-0,293
1-10	-0,030	-0,056	-0,173	-0,182	-1,133	-0,554	0,000	-0,001	-1,964	-0,132	-0,094	-4,751	-4,286	-4,909
1-11	-0,026	-0,048	-0,152	-0,159	-1,190	-0,600	-0,001	0,000	-2,083	-0,167	-0,124	-5,208	-4,714	-5,392
2-0	-2,667	-4,000	-6,667	-6,769	-1,000	-2,133	-5,594	-5,762	-1,333	-4,762	-4,964	-1,818	-2,004	-1,833
2-1	-1,333	-2,133	-4,103	-4,190	-0,133	-0,667	-3,126	-3,267	-0,190	-2,400	-2,564	-0,290	-0,381	-0,293
2-2	-0,800	-1,333	-2,857	-2,933	0,000	-0,190	-1,964	-2,083	0,000	-1,333	-1,466	0,000	-0,008	0,000
2-10	-0,103	-0,190	-0,606	-0,638	-0,762	-0,267	-0,132	-0,167	-1,333	0,000	-0,004	-3,333	-2,926	-3,451
2-11	-0,088	-0,164	-0,535	-0,564	-0,831	-0,314	-0,094	-0,124	-1,466	-0,004	0,000	-3,776	-3,337	-3,916
10-0	-3,636	-6,667	-20,000	-20,952	-1,667	-4,103	-17,316	-18,333	-2,857	-15,152	-16,187	-6,667	-7,435	-6,875
10-1	-3,030	-5,594	-17,316	-18,182	-1,133	-3,126	-14,727	-15,653	-1,964	-12,653	-13,594	-4,751	-5,411	-4,909
10-2	-2,564	-4,762	-15,152	-15,942	-0,762	-2,400	-12,653	-13,500	-1,333	-10,667	-11,524	-3,333	-3,896	-3,451
10-10	-0,952	-1,818	-6,667	-7,097	0,000	-0,290	-4,751	-5,208	0,000	-3,333	-3,776	0,000	-0,023	0,000
10-11	-0,866	-1,656	-6,144	-6,548	-0,004	-0,214	-4,286	-4,714	-0,008	-2,926	-3,337	-0,023	0,000	-0,024
11-0	-3,667	-6,769	-20,952	-22,000	-1,692	-4,190	-18,182	-19,290	-2,933	-15,942	-17,064	-7,097	-7,923	-7,333
11-1	-3,103	-5,762	-18,333	-19,290	-1,190	-3,267	-15,653	-16,667	-2,083	-13,500	-14,524	-5,208	-5,926	-5,392
11-2	-2,659	-4,964	-16,187	-17,064	-0,831	-2,564	-13,594	-14,524	-1,466	-11,524	-12,462	-3,776	-4,396	-3,916
11-10	-1,048	-2,004	-7,435	-7,923	-0,004	-0,381	-5,411	-5,926	-0,008	-3,896	-4,396	-0,023	-0,095	-0,024
11-11	-0,957	-1,833	-6,875	-7,333	0,000	-0,293	-4,909	-5,392	0,000	-3,451	-3,916	0,000	-0,024	0,000

On vérifie que la fusion de lignes ayant des proportions identiques correspond à des DeltaKhi2 nuls (fusions de lignes 0-n et 0-n', ou n-n et n'-n').

Outre les fusions de valeur nulle, les fusions préférées pour une ligne de Khi2 0-1 sont dans l'ordre :

1-11 : -0,026
1-10 : -0,030
2-11 : -0,088
2-10 : -0,103
1-2 : -0,333
1-1: -0,667
2-2 : -0,800
10-11 : -0,866
10-10 : -0,952
11-11 : -0,957

Les fusions préférées (de valeur non nulle) pour une ligne de Khi2 10-11 sont dans l'ordre :

1-1 : -0,004
2-2 : -0,008
10-10 : -0,023
11-11 : -0,024
11-10 : -0,095
1-2 : -0,214
2-1 : -0,381
0-1 : -0,866

L'ordre des fusions induit par le critère du DeltaKhi2 correspond bien à l'intuition.

Pour une valeur de Khi2/(Ndl+1) >= 1 (ce qui est le cas souvent dans les étapes de discrétisations), toutes les fusions précédentes seraient acceptées dans la procédure Khiops, car elles ont un DeltaKhi2 d'amplitude inférieure à 1. Par exemple, la ligne 10-11 pourrait être fusionnée avec la ligne 1-2 (DeltaKhi2=-0,214), avec 2-1 (DeltaKhi2 = -0,381), avec 0-1 (DeltaKhi2 = -0,866), mais pas avec la 1-0 (DeltaKhi2 = -1,048). Ce dernier seuil est inférieur à 1,05 et serait donc accepté pour des valeurs de Khi2 de l'ordre de 1,1(Ndl+1).

Dans le cas de la fusion d’une ligne élémentaire 0-1 avec une ligne de cardinal n et de proportions identiques aux proportions des modalités cibles, DeltaKhi2 = -n/(n+1), donc DeltaKhi2 est d’amplitude inférieure à 1. Une ligne élémentaire 0-1 sera toujours intéressante à fusionner avec un intervalle de proportions identiques (ou similaires) aux proportions cibles, et ce d’autant plus que le cardinal de l’intervalle est faible.

Dans le cas de la fusion de deux lignes de même cardinal n, pour p = 0,5, la fusion entre les deux lignes sera acceptée si la différence des proportions entre les lignes est bornée par l’inverse de la racine de n, soit plus précisément si $\left|a-b\right| \leq \frac{\sqrt{D/2}}{\sqrt{n}}$.

En se ramenant aux effectifs et en posant a = k/n et b=k’/n, la fusion entre deux lignes de même cardinal est acceptée si la différence entre les effectifs observés est inférieure à la racine de n, ou plus exactement si $\left|k-k'\right| \leq \sqrt{D/2}\sqrt{n}$.

6.8 Conclusion

La méthode Khiops utilise la loi de probabilité du Khi2 sur de très grandes plages de valeurs de ses paramètres. Les méthodes d’approximation habituelles du Khi2 n’étant pas utilisables, nous avons utilisé une approximation du logarithme de la probabilité du Khi2 pour s’affranchir des problèmes numériques. Pour l’évaluation du DeltaKhi2 qui est sensible aux variations fines de la loi du Khi2, nous avons utilisé une nouvelle méthode d’approximation spécialement conçue pour l’algorithme Khiops. Quelques résultats théoriques, notamment concernant les bornes et le comportement asymptotique du DeltaKhi2 nous ont permis de qualifier le comportement de cette fonction. Les évaluations numériques de l’approximation nous ont permis de confirmer sa fiabilité sur de très grandes plages de valeur.