france telecom R&D

Note Technique
NT/FTR&D/7339

26 septembre 2001

Khiops: discrétisation des attributs numériques pour le Data Mining

Marc Boullé (DTL/DLI)

Vu, pour accord le directeur de DTL

JM. Pitié

Vu, le chef du département DLI

JF. Cloarec

Date : 26 septembre 2001

Résumé : Dans le domaine de l'apprentissage supervisé, certains modèles sont adaptés uniquement aux données qualitatives. Ces modèles procèdent alors à une étape de discrétisation des attributs numériques pour pouvoir les prendre en compte. De nombreuses méthodes de discrétisation ont été proposées dans la bibliographie, qui se basent sur des critères statistiques, informationnels ou encore d'autres critères dédiés. Nous proposons ici une nouvelle méthode de discrétisation, Khiops, basée sur la statistique du Khi2. Contrairement aux méthodes de discrétisation apparentées ChiMerge et ChiSplit, cette méthode optimise le critère du Khi2 globalement sur l'ensemble du domaine de discrétisation et ne nécessite aucun paramétrage de critère d'arrêt de la discrétisation. Une étude théorique complétée par des expérimentations montre la robustesse de la méthode et la qualité prédictive des discrétisations obtenues.

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

France Télécom R&D
Branche Développement
38-40 rue du Général Leclerc
92794 Issy-les Moulineaux Cedex9
France
Téléphone : 01 45 29 44 44
Téléphone international : +33 1 45 29 44 44 44

TABLE DES MATIERES

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 (Zighed 1996) 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 de l'intervalle complet à discrétiser et cherchent le meilleur point de coupure de l'intervalle en optimisant le critère choisi. La méthode est appliquée itérativement aux deux sous intervalles jusqu'à ce qu'un critère d'arrêt soit rencontré.
Les méthodes ascendantes partent d'intervalles élémentaires et cherchent la meilleure fusion de deux intervalles adjacents en optimisant le critère choisi. La méthode est appliquée itérativement aux intervalles restant jusqu'à ce qu'un critère d'arrêt soit rencontré.
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.

La méthode de discrétisation Khiops est 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 décroit 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. Nous montrerons qu'en dépit de cette approche globale, l'algorithme associé à la méthode Khiops est en Nlog(N) ou N est le nombre d'individus à discrétiser.

Cette complexité algorithmique est la même que pour l'algorithme ChiMerge optimisé. Nous comparerons la méthode Khiops avec d'autres méthodes de discrétisation et procéderons à des expérimentations. Enfin, nous étudierons les problèmes numériques liés au calcul de la loi du Khi2 pour un paramétrage extrême (très grand nombre de degrés de liberté et très grande valeur du Khi2).

Le document est organisé de la façon suivante.
La partie 1 rappelle les principes du test du Khi2 et étudie quelques unes de ses propriétés. La partie 2 présente l'algorithme Khiops 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 procède à des expérimentations. L'annexe étudie les problèmes de sensibilité numérique liés à l'approximation de la loi du Khi2.

1. LE TEST D'INDEPENDANCE DU KHI2 : PRINCIPES

Soit S une variable source, et T une variable cible.
On cherche à savoir si les variables S et T sont indépendantes.
On construit le tableau de contingence, comptant le nombre d'individus pour chaque couple de valeurs de S et T. Dans le tableau 1 par exemple, le couple de valeur (d, B) a été observé 10 fois.

S\T	A	B	C
a	0	2	1
b	2	2	2
c	3	8	0
d	5	10	2
e	8	9	1

Tableau 1 : Exemple de tableau de contingence

Le test du Khi2 permet de tester l'hypothèse d'indépendance des deux lois.
Le Khi2 est calculé à partir du tableau de contingence.

S\T	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

Tableau 2 : Tableau de contingence utilisé pour le calcul du Khi2

n_{ij} : Nombre d'individus pour la $i^{ème}$ valeur de la variable S et la $j^{ème}$ valeur de la variable T
$n_{i.}$: Nombre total d'individus pour la $i^{ème}$ valeur de la variable S
$n_{.j}$: Nombre total d'individus pour la $j^{ème}$ valeur de la variable T
N : Nombre total d'individus
I : Nombre de modalités de la variables T (ici 3)
J : Nombre de modalités de la variable S (ici 5)

Soit $e_{ij} = n_{i.} * n_{.j} / N$.
e_{ij} représente le nombre d'individus de la case (i, j) si les lois étaient indépendantes. e_{ij} est l'effectif théorique de la case (i,j).
La valeur du Khi2 est une mesure sur l'ensemble du tableau de l'écart entre les nombres d'individus observés (effectif observé) et les nombres d'individus théoriques (effectif théorique) si les lois étaient indépendantes. La valeur du Khi2 est donc une mesure de l'écart à l'hypothèse d'indépendance des variables.

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

Sous l'hypothèse nulle d'indépendance, la valeur du Khi2 suit une loi du Khi2 à (I-1)*(J-1) degrés de liberté, ce qui permet de construire un test rejetant l'hypothèse quand la valeur du Khi2 est suffisamment grande. Plus la valeur du Khi2 est importante, moins l'hypothèse d'indépendance des variables est probable.
Par abus de langage, on parlera dans la suite de probabilité d'indépendance des variables.

Proba Degrés	0,99	0,98	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,0002	0,0008	0,0039	0,0158	0,06	0,15	0,45	1,07	1,64	2,71	3,84	5,41	6,63	10,83
2	0,02	0,04	0,10	0,21	0,45	0,71	1,39	2,41	3,22	4,61	5,99	7,82	9,21	13,82
3	0,11	0,18	0,35	0,58	1,01	1,42	2,37	3,66	4,64	6,25	7,81	9,84	11,34	16,27
4	0,30	0,43	0,71	1,06	1,65	2,19	3,36	4,88	5,99	7,78	9,49	11,67	13,28	18,47
5	0,55	0,75	1,15	1,61	2,34	3,00	4,35	6,06	7,29	9,24	11,07	13,39	15,09	20,52
6	0,87	1,13	1,64	2,20	3,07	3,83	5,35	7,23	8,56	10,64	12,59	15,03	16,81	22,46
7	1,24	1,56	2,17	2,83	3,82	4,67	6,35	8,38	9,80	12,02	14,07	16,62	18,48	24,32
8	1,65	2,03	2,73	3,49	4,59	5,53	7,34	9,52	11,03	13,36	15,51	18,17	20,09	26,12
9	2,09	2,53	3,33	4,17	5,38	6,39	8,34	10,66	12,24	14,68	16,92	19,68	21,67	27,88
10	2,56	3,06	3,94	4,87	6,18	7,27	9,34	11,78	13,44	15,99	18,31	21,16	23,21	29,59

Tableau 3 : Table du Khi2 entre 1 et 10 degrés de liberté.

Par exemple, pour un tableau de contingence de dimension 5*3, le nombre de degrés de liberté de la loi du Khi2 associée est 8. Si on trouve une valeur de Khi2 de 20, cela signifie que l'hypothèse d'indépendance des variables à une probabilité d'environ 1%. Il est donc raisonnable de rejeter l'hypothèse d'indépendance dans ce cas.

Sensibilité à l'indépendance des variables

Plus les variables sont indépendantes, plus les effectifs observés sont proches des effectifs théoriques. Dans ce cas la valeur du Khi2 est faible et la probabilité d'indépendance est donc forte.

Sensibilité aux effectifs

Si on multiplie tous les effectifs par un facteur constant k en gardant les mêmes proportions de modalités cibles, la nouvelle valeur du Khi2 est multipliée par k.

$$NewKhi2 = \sum_i \sum_j \frac{(k\,n_{ij} - k\,e_{ij})^2}{k\,e_{ij}} = k.Khi2$$

La probabilité d'indépendance diminue alors très rapidement avec la taille de la population. Cette propriété provient du caractère statistique du test. Avec une petite population, une distribution non homogène est relativement fréquente, mais est peu fiable pour rejeter l'hypothèse d'indépendance des variables. Pour une population plus grande, il devient de moins en moins probable qu'une non-homogénéité de la distribution soit due au hasard.

Sensibilité aux proportions observées de modalités cibles

Soit une distribution des modalités cible $p_1, p_2, \ldots p_J$. $\sum_j p_j = 1$

Etudions l'influence d'une ligne de Khi2, d'effectif n, pour des proportions sur la ligne de modalités cibles a_j. $\sum_j a_j = 1$

Les effectifs observés et théoriques de la ligne de Khi2 sont $a_j n$ et $p_j n$.

La contribution de la ligne au Khi2 est donc

$$Khi2l = \sum_j \frac{(n(p_j - a_j))^2}{p_j n}$$

$$Khi2l = n\sum_j \frac{(p_j - a_j)^2}{p_j} = n\left(\sum_j \frac{a_j^{\,2}}{p_j} - 1\right)$$

Le Khi2 ligne est proportionnel à l'effectif ligne, et varie comme une combinaison des carrés des écarts entre les proportions théoriques et observées de modalités cibles, pondérées par l'inverse des proportions cibles. L'écart aux modalités cibles de faibles proportions est donc favorisé.

Sensibilité de la loi du Khi2

On a vu que la valeur du Khi2 varie de façon linéaire avec les effectifs, et approximativement de façon quadratique avec les écarts entre les proportions observées et théoriques. Par contre, en se basant sur la

table du Khi2 du tableau 3, on observe que la probabilité d'indépendance varie de façon exponentielle avec la valeur du Khi2. Des variations faibles de la valeur du Khi2 entraînent des variations importantes de la probabilité d'indépendance correspondante.

2. METHODE DE DISCRETISATION KHIOPS

2.1. Algorithme

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)

2.2. 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 ainsi redéfini 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 pour assurer à la fois la fiabilité statistique du test du Khi2 et prévenir les problèmes de sur-apprentissage.

2.3. 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 le suivant:

Valeur Sepal width	Iris versicolor	Iris Virginica	Iris setosa	Total
2	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	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	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

Intervalle fusionné	Khi2 Résultant
] -∞ ; 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

Tableau 4 : Table de contingence pour l'attribut sepal width de la base Iris. Evaluation des fusions.

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 5 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.

Valeur Sepal width	Iris versicolor	Iris virginica	Iris setosa	Total
2	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	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	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

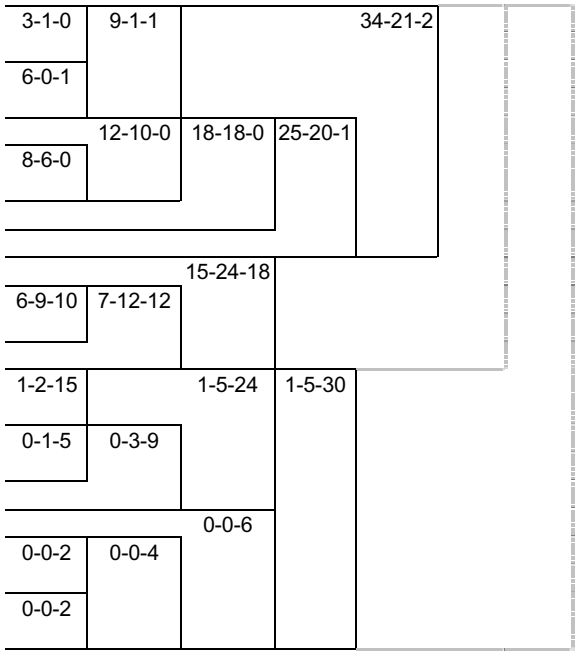

Tableau 5 : Fusions successives des intervalles pour arriver à une discrétisation en trois intervalles

Au bout d'une vingtaine d'étapes, on arrive à la loi discrétisée suivante:

Valeur 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

Intervalle Fusionné	Khi2 Résultant
] -∞ ; 3,35]	54,17
]2,95 ; +∞]	43,97

Tableau 6 : Table de contingence pour l'attribut sepal width discrétisé de la base Iris

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.

2.4. Complexité algorithmique

On va évaluer la complexité algorithmique de la méthode de discrétisation Khiops par rapport au nombre d'individus N de la base de données de travail. Dans le pire des cas, les individus 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/supressions 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 (de 100000 à 1000000 d'individus).

2.5. Propriétés de la fusion des lignes de Khi2

Soit une distribution des modalités cible p_1, p_2, …p_J. $\sum_j p_j = 1$

Soit une première ligne de Khi2, d'effectif n, pour des proportions de modalités cibles a_j. $\sum_j a_j = 1$

Soit une seconde ligne de Khi2, d'effectif n', pour des proportions de modalités cibles b_j. $\sum_j b_j = 1$

Les effectifs observés et théoriques de la première ligne de Khi2 sont $a_j n$ et $p_j n$.
Les effectifs observés et théoriques de la seconde ligne de Khi2 sont $b_j n'$ et $p_j n'$.

Les Khi2 lignes sont $Khi2l = n\left(\sum_j \frac{a_j^{\,2}}{p_j} - 1\right)$ et $Khi2l' = n'\left(\sum_j \frac{b_j^{\,2}}{p_j} - 1\right)$.

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

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

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

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

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.
Si les deux lignes ont exactement les mêmes proportions de modalités cibles ($a_j = b_j$), alors la fusion de ces deux lignes ne fait pas diminuer le Khi2. La fusion de deux lignes aux proportions identiques (ou très proches) diminue donc la valeur de la probabilité d'indépendance. Pour diminuer la probabilité d'indépendance, il est plus important d'être similaire pour les petits p_j que pour les grands p_j.
Pour un rapport d'effectifs constant, la décroissance du Khi2 est proportionnelle à l'effectif global des deux lignes. Les fusions avec effectifs faibles ont plus de chance de diminuer la probabilité d'indépendance
Pour un effectif global des deux lignes constant, la décroissance du Khi2 est maximale quand les effectifs des deux lignes sont identiques. Les fusions avec effectifs différents diminue donc davantage la probabilité d'indépendance.

En résumé, la probabilité d'indépendance diminue (le Khi2 décroît le moins) selon les facteurs suivants :

- faibles effectifs sur les lignes
- effectifs différents entre les lignes
- proportions similaires entre les lignes (surtout pour les petits p_j)

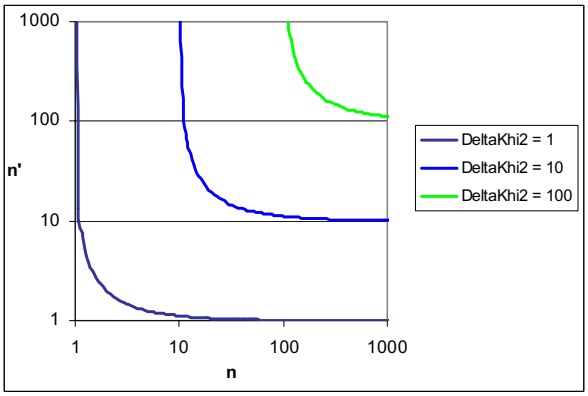

Figure 1 : Influence des effectifs

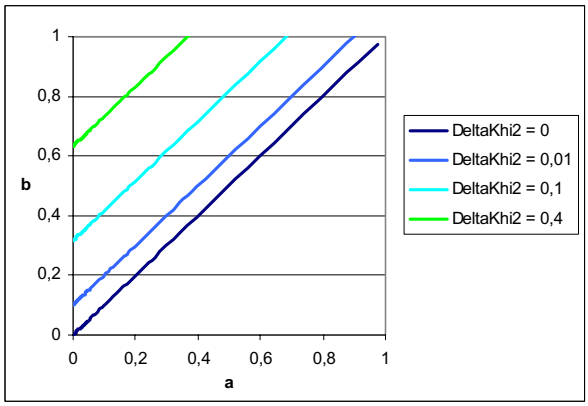

Figure 2 : Influence des proportions

Remarque
Les fusions de lignes ayant les même proportions de modalités cibles sont optimales pour l'algorithme et seront donc effectuées les premières (aux contraintes d'effectif minimum près). Toutes les valeurs adjacentes ayant même modalité cible peuvent donc être regroupées pour constituer les intervalles initiaux lors de la phase d'initialisation de l'algorithme.

2.6. De la méthode à son implémentation

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.

Le calcul de l'effectif minimal théorique doit tenir compte plus précisément des lois cibles à plusieurs modalités pour forcer le regroupement d'intervalles qui ne paraissent dissemblables que par le hasard de l'échantillon. Ce calcul n'a ici été fait qu'approximativement.

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.

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 proche de zéro.

Les limites de la méthode proviennent d'avantage de son implémentation que de son principe. Le problème le plus critique est celui de l'évaluation de la loi du Khi2.
Nous montrerons que l'approximation de l'effectif minimal et l'heuristique gloutonne utilisée permettent d'obtenir des résultats de très bonne qualité avec des temps de calcul très rapides.
Nous étudieront également en annexe de nouvelles 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.

3. COMPARAISON THEORIQUE AVEC LES METHODES BASEES SUR LE KHI2

3.1. Comparaison avec ChiMerge

Soit une distribution des modalités cible $p_1, p_2, \ldots p_J$. $\sum_j p_j = 1$

Soit une première ligne de Khi2, d'effectif n, pour des proportions de modalités cibles a_j. $\sum_j a_j = 1$

Soit une seconde ligne de Khi2, d'effectif n', pour des proportions de modalités cibles b_j. $\sum_j b_j = 1$

Les effectifs observés et théoriques de la première ligne de Khi2 sont $a_j n$ et $p_j n$.
Les effectifs observés et théoriques de la seconde ligne de Khi2 sont $b_j n'$ et $p_j n'$.

Les Khi2 lignes sont $Khi2l = n\left(\sum_j \frac{a_j^2}{p_j} - 1\right)$ et $Khi2l'' = n'\left(\sum_j \frac{b_j^2}{p_j} - 1\right)$.

On a vu que pour la méthode Khiops, le calcul du DeltaKhi2 résultant de la fusion de deux lignes conduit à :

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

Pour la méthode ChiMerge, on considère le tableau du Khi2 local aux deux lignes. Dans ce contexte local, la distribution des modalités cibles $q_1, q_2, \ldots q_J$ a pour valeurs $q_j = \frac{a_j n + b_j n'}{n+n'}$.

Pour évaluer l'intérêt de la fusion des deux lignes, on calcule le Khi2 de cette table locale du Khi2.

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

$$SommeKhi2l = \left(\sum_j \frac{\left(a_j^2 n + b_j^2 n'\right)(n+n')}{(a_j n + b_j n')}\right) - (n+n')$$

$$SommeKhi2l = \left(\sum_j \frac{(a_j n + b_j n')^2 + nn'(a_j - b_j)^2}{(a_j n + b_j n')}\right) - (n+n')$$

$$SommeKhi2l = \left(nn'\sum_j \frac{(a_j - b_j)^2}{(a_j n + b_j n')}\right) + \left(\sum_j a_j n + b_j n'\right) - (n+n')$$

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

Le calcul du critère d'arrêt pour les méthodes Khiops et ChiMerge conduit donc a une expression mathématique identique. L'interprétation du critère est radicalement différente. La distribution des modalités cibles est globale a 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 fondamentale entre les deux 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.

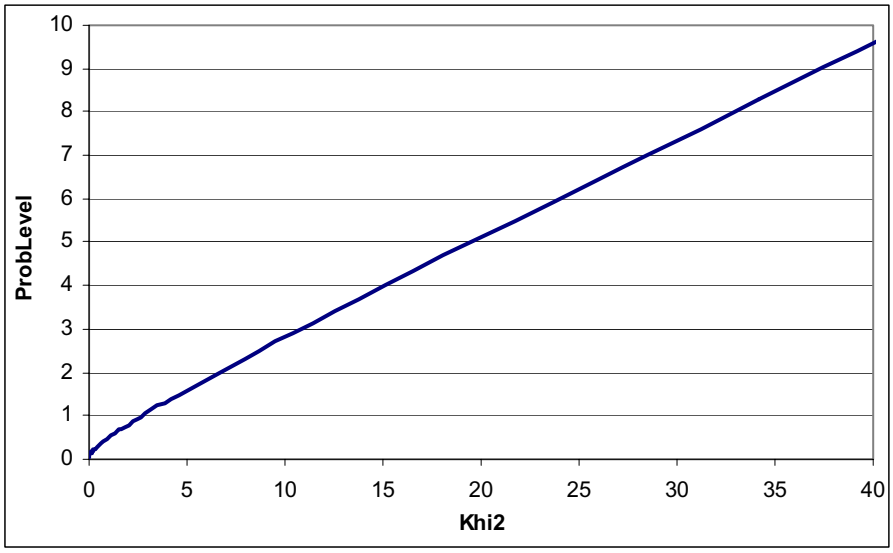

Figure 3 : Loi du Khi2 à un degré de liberté

On va prendre l'exemple de la fusion des deux lignes de même effectif ayant des proportions de modalités cibles légèrement différentes pour une loi cible à deux modalités équiréparties.

n(p-e)	n(1-p+e)	n
n(p+e)	n(1-p-e)	n
2np	2n(1-p)	2n

Les probabilités locales au tableau du Khi2 des deux lignes sont p et 1-p. La différence d'effectifs observés entre les deux lignes est D=2ne pour une même modalité cible

Dans ce cas, on a $DeltaKhi2 = -8ne^2 = -2\frac{D^2}{n}$ et $SommeKhi2l = \frac{2ne^2}{p(1-p)} = \frac{D^2}{2np(1-p)}$.

Les seuils d'acceptation usuels du DeltaKhi2 varie de 1 à 10 quand Khi2/(Ndl+1) varie de 1 à 100 (cf. annexe). Cela signifie que pour l'algorithme Khiops, la fusion de deux lignes est acceptée dès que la différence des effectifs observés entre les deux lignes est au maximum de l'ordre de racine de n. Cette différence augmente avec le Khi2 global. Pour avoir un comportement équivalent dans ChiMerge, il faut que p=0,5. Dans ce cas, il faut fixer un seuil de Khi2 entre 0,3 (Khi2 = 1) et 0,001 (Khi2 = 10). Par ailleurs, l'algorithme du ChiMerge est très sensible aux probabilités cibles p locales aux deux lignes. Si l'on passe à p=0,1 au lieu de 0,5, le calcul SommeKhi2l donne un résultat presque trois fois supérieur. Pour un résultat équivalent, il faut alors ajuster leur seuil de Khi2 entre 0,1 (Khi2 = 3) et 0,0000001 (Khi2 = 30). Le comportement « intuitif » de la méthode Khiops qui consiste à autoriser toute fusion de deux lignes « similaires » n'est pas reproductible avec l'algorithme ChiMerge.

On va prendre la table suivante pour illustrer la difficulté de choisir un seuil de Khi2 pour l'algorithme ChiMerge :

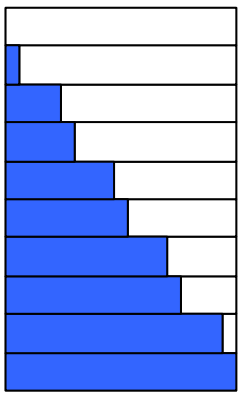

Table initiale		Khiops	ChiMerge		Table finale	
		DeltaKhi2l	SommeKhi2l	Seuil		
0	100					
6	94	-0,72	6,19	0,013	6	194
24	76	-6,48	12,71	0,000		
30	70	-0,72	0,91	0,339	54	146
47	53	-5,78	6,10	0,013		
53	47	-0,72	0,72	0,396	100	100
70	30	-5,78	6,10	0,013		
76	24	-0,72	0,91	0,339	146	54
94	6	-6,48	12,71	0,000		
100	0	-0,72	6,19	0,013	194	6

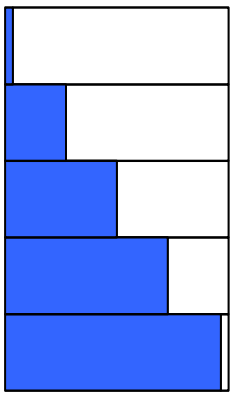

Tableau 7 : Choix de la meilleure fusion d'intervalle pour Khiops et ChiMerge

On a ici un Khi2 total pour la table globale de 449,2 égale à environ 50 fois le nombre de degrés de liberté. En se référant à la table 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 fait ici, en choisissant le seuil à 0,01 (pour accepter la fusion entre 0-100 et 6-94), l'algorithme ChiMerge va donner un résultat cohérent : les premières fusions vont en effet avoir pour effet d'interdire la fusion entre les lignes 30-70 (fusionnée avec 24-76) et 47-53 (fusionnée avec 53-47). Les deux méthodes aboutissent dans ce cas à la même table finale.

Dans l'exemple suivant (Khi2 total=378), seul l'algorithme Khiops permet d'aboutir aux fusions naturelles. Pour ChiMerge, la fusion entre les lignes 33-67 et 50-50 est préférée à la fusion « naturelle » entre les lignes 0-100 et 6-94 quel que soit le seuil choisi.

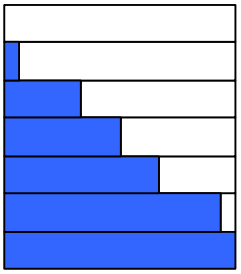

Table initiale		Khiops	ChiMerge		Table finale	
		DeltaKhi2l	SommeKhi2l	Seuil		
0	100					
6	94	-0,72	6,19	0,013	6	194
33	67	-14,58	23,22	0,000	33	67
50	50	-5,78	5,95	0,015	50	50
67	33	-5,78	5,95	0,015	67	33
94	6	-14,58	23,22	0,000		
100	0	-0,72	6,19	0,013	194	6

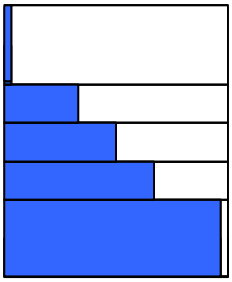

Tableau 8 : Choix de la meilleure fusion d'intervalle pour Khiops et ChiMerge

On va utiliser le dernier exemple suivant (Khi2 total=3800) pour illustrer la prise en compte des facteurs d'échelle. Dans la méthode Khiops, on se trouve ici à un niveau de Khi2/(Ndl+1) supérieur à 600. A ce niveau de Khi2, les fusions "naturelles" entre 0-1000 et 50-950 sont acceptées, les autres sont largement écartées. Pour ChiMerge, la fusion entre 350-650 et 500-500 est systématiquement préférée à la fusion entre 0-1000 et 50-950 quel que soit le seuil choisi. Il faut noter que ce seuil pour ChiMerge est ici de l'ordre de 10^{-12}. Il dépend donc fortement des effectifs en jeu, ce qui rend son ajustement manuel extrêmement délicat.

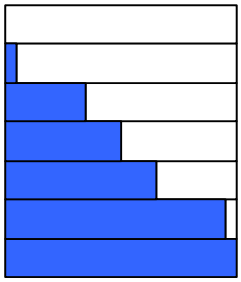

Table initiale	
0	1000
50	950
350	650
500	500
650	350
950	50
1000	0

Khiops
DeltaKhi2l
-5
-180
-45
-45
-180
-5

ChiMerge	
SommeKhi2l	Seuil
51,28	8,00E-13
281,25	4,01E-63
46,04	1,16E-11
46,04	1,16E-11
281,25	4,01E-63
51,28	8,00E-13

Table finale	
50	1950
350	650
500	500
650	350
1950	50

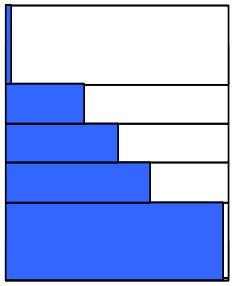

Tableau 9 : Choix de la meilleure fusion d'intervalle pour Khiops et ChiMerge

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 évaluées équitablement parmi l'ensemble de toutes les fusions possibles.

3.2. 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.

Prenons l'exemple d'une loi cible à deux modalités équiréparties, pour laquelle seule une ligne de la table du Khi2 présente des proportions de modalités cibles différentes des proportions globales.

np	n(1-p)	n
np	n(1-p)	n
…	…	…
np	n(1-p)	n
n(p-e)	n(1-p+e)	n
np	n(1-p)	n
…	…	…
np	n(1-p)	n
np	n(1-p)	n
n(Ip-e)	n(I-Ip+e)	In

On va calculer le critère du ChiSplit pour un premier intervalle constitué de i lignes de type np-n(1-p) et un second intervalle contenant le reste de la table.

$$Khi2Split_i = \frac{(nip - ni(p - e/I))^2}{ni(p - e/I)} + \frac{(ni(1-p) - ni(1-p+e/I))^2}{ni(1-p+e/I)} +$$

$$\frac{(n((I-i)p - e) - n(I-i)(p - e/I))^2}{n(I-i)(p - e/I)} + \frac{(n((I-i)(1-p) + e) - n(I-i)(1-p+e/I))^2}{n(I-i)(1-p+e/I)}$$

$$\boxed{Khi2Split_i = \frac{(e/I)^2}{(p - e/I)(1-p+e/I)}\left(1 + \left(\frac{i}{I-i}\right)^2\right)}$$

Si la ligne singulière est en i_0 avec $i_0 <= I/2$, la valeur du critère croît jusqu'à i_0 puis décroît ensuite (la ligne singulière étant passée de l'autre côté du point de coupure, il faut utiliser la formule avec I-i). La

coupure se fait juste au ras de la ligne singulière, qui sera isolée en ré-appliquant l'algorithme sur le sous intervalle comportant cette ligne. La méthode ChiSplit arrive donc à isoler correctement la singularité.
Le calcul du ChiSplit met néanmoins en lumière ses problèmes d'utilisation. Le critère d'arrêt est très délicat à ajuster car il dépend de facteurs d'échelle (nombre de lignes du tableau), de l'importance des singularités à détecter, et de la position de la singularité dans la table du Khi2. En effet, la valeur de Khi2Split (maximale au point de coupure) varie du simple (singularité en i_0=1) au double (singularité en i_0=I/2) selon la position de la singularité, ce qui rend un ajustement optimal impossible dans le cas de plusieurs singularités présentes à des positions différentes.

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

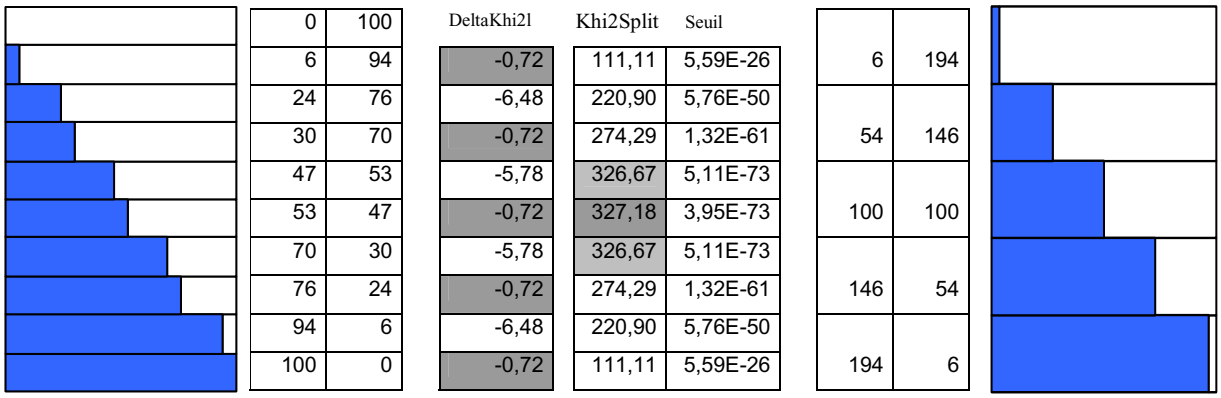

Table initiale		Khiops	ChiSplit		Table finale	
		DeltaKhi2l	Khi2Split	Seuil		
0	100					
6	94	-0,72	111,11	5,59E-26	6	194
24	76	-6,48	220,90	5,76E-50		
30	70	-0,72	274,29	1,32E-61	54	146
47	53	-5,78	326,67	5,11E-73		
53	47	-0,72	327,18	3,95E-73	100	100
70	30	-5,78	326,67	5,11E-73		
76	24	-0,72	274,29	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

Tableau 10 : Choix de la meilleure fusion d'intervalle pour Khiops et ChiSplit

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), 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 constitue donc une faiblesse intrinsèque pour l'identification des régularités locales de la variable à discrétiser.

4. EXPERIMENTATIONS

4.1. Description des expérimentations menées

Nous avons effectué une série de test sur des jeux d'essai théoriques parfaitement connus, à savoir le mélange de deux classes suivant chacune une loi de distribution gaussienne. L'objectif est d'étudier le comportement de la méthode de discrétisation Khiops en fonction de la taille de l'échantillon et du degré de séparabilité des deux classes, ajustable par l'écart type des gaussiennes.

Les jeux d'essai sont constitués de la façon suivante :

- Chaque jeu d'essai correspond à un échantillon paramétré un écart type *ET* et une taille d'échantillon *Taille*.
- Chaque individu est représenté par une variable continue "Value" et une variable cible "Class" à prédire.
- La variable à prédire Class a deux valeurs 0 et 1 équiréparties.
- La loi de distribution des 0 est une gaussienne de moyenne 0 et d'écart type *ET*.
- La loi de distribution des 1 est une gaussienne de moyenne 1 et d'écart type *ET*.

On étudie la discrétisation de l'attribut Value pour l'attribut à prédire Class. Afin d'obtenir des valeurs statistiquement fiables, l'expérimentation est répétée 100 fois pour chaque couple de valeur (*ET*, *Taille*). Les écarts types étudiés sont 0,1, 0,2, 0,25, 0,3, 0,4, 0,5, 0,6, 1, 2, 10, ce qui permet de passer progressivement de classes presque parfaitement séparables à des classes pratiquement mélangées aléatoirement. Les tailles d'échantillon étudiées sont 100, 1000, 10000, 100000, 1000000, ce qui permet de passer d'échantillons peu fiables statistiquement à des échantillons très volumineux, correspondant surtout à des tests de volumétrie.

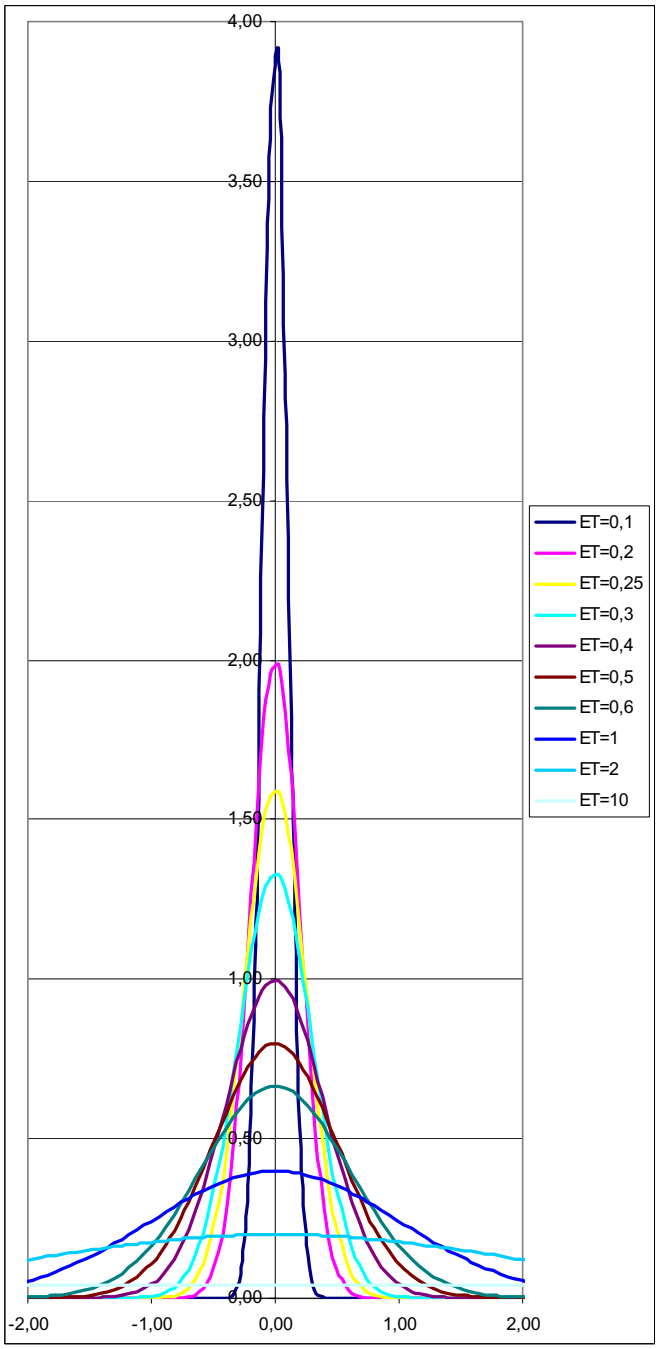

Figure 4: Lois gaussiennes utilisées pour l'expérimentation

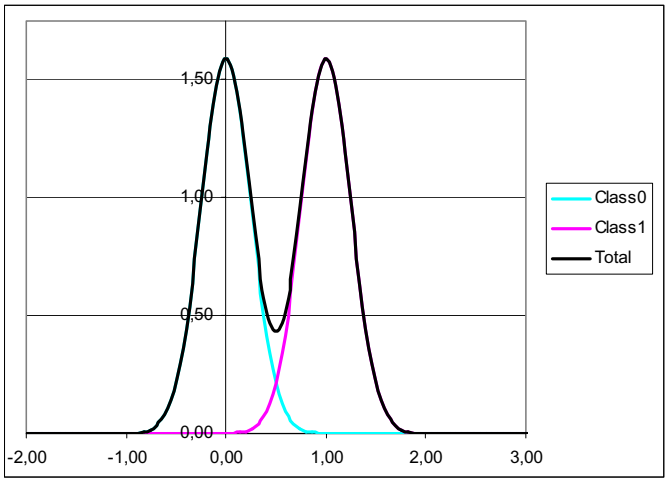

Figure 5 : Mélange de deux gaussiennes d'écart type 0,25

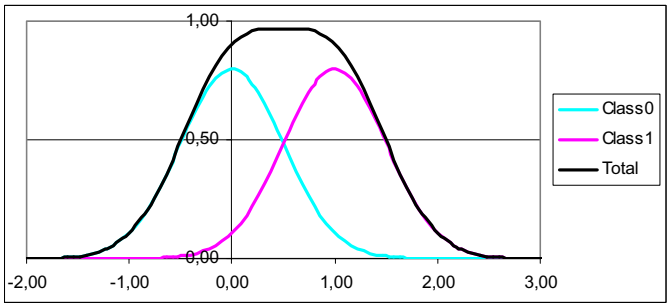

Figure 6 : Mélange de deux gaussiennes d'écart type 0,5

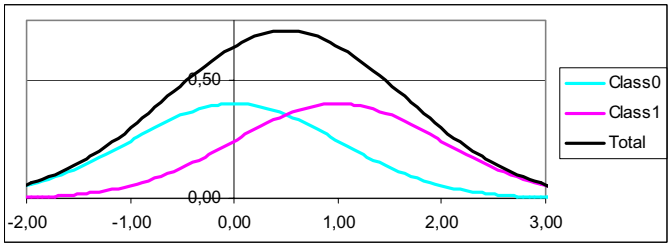

Figure 7 : Mélange de deux gaussiennes d'écart type 1

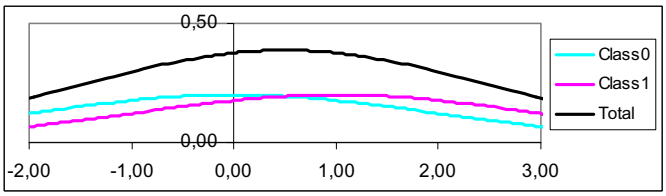

Figure 8 : Mélange de deux gaussiennes d'écart type 2

Une autre façon de présenter le problème du mélange des gaussiennes est de visualiser la proportion de la classe 0 (par exemple) en fonction de la valeur de la variable continue à discrétiser. Les classes sont équiréparties pour la valeur 0,5, et la transition entre la zone où la classe 0 est majoritaire et celle où la classe 1 est majoritaire est d'autant plus rapide que l'écart type des gaussienne est faible.

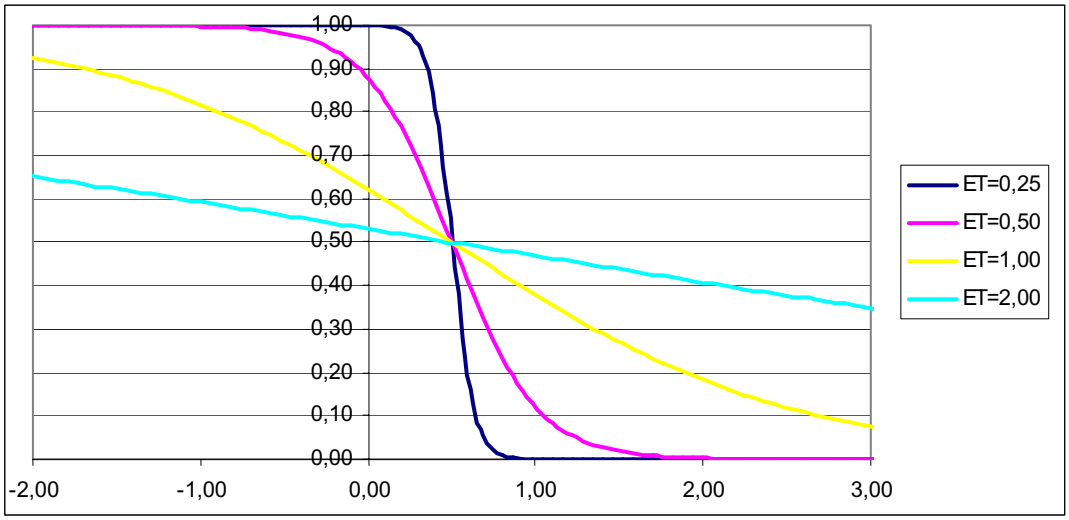

Figure 9 : Proportion de la classe 0 pour différents écarts types des gaussiennes

Pour illustrer la variabilité statistique des échantillons, on va visualiser la proportion de la classe 0 sur des histogrammes constitués de partiles de taille 25 et 100 pour la séparation de deux gaussiennes d'écart type 1 sur un échantillon de taille 1000. Ces histogrammes constituent des discrétisations non supervisées et montrent l'impact du choix des effectifs par intervalle. On voit clairement qu'une taille d'effectif par intervalle trop petite conduit à refléter trop fidélement les aléas de l'échantillon, alors qu'une taille trop importante (à la limite deux intervalles uniquement) conduirait à approximer trop grossièrement la courbe de répartition réelle.

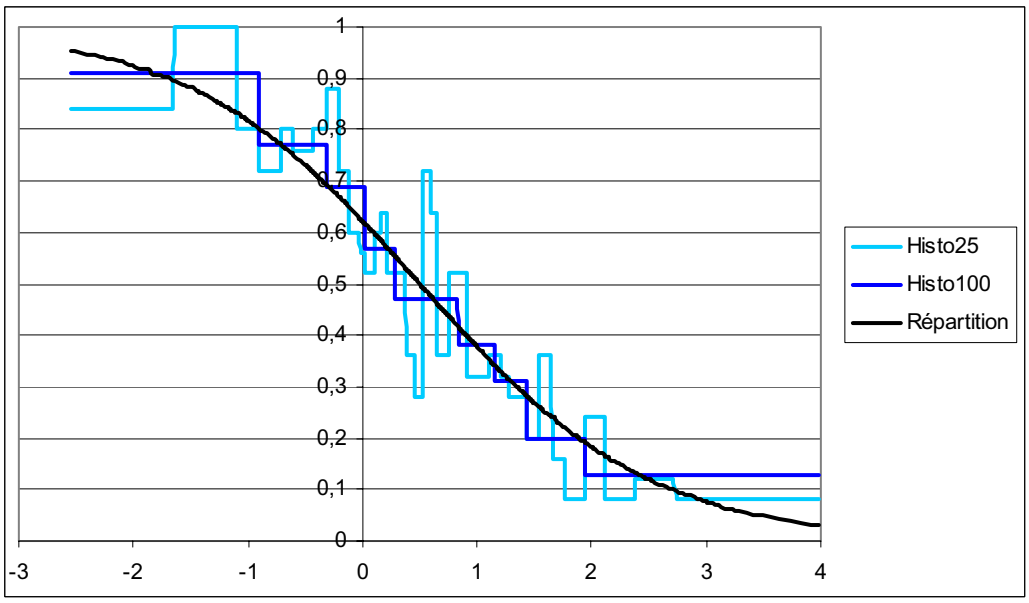

Figure 10 : Proportion de la classe 0 et histogrammes par partiles pour deux gaussiennes d'écart type 1 sur un échantillon de taille 1000

Pour chaque jeu de paramètres, on a mesuré les indicateurs suivants :

- Temps de discrétisation
- Indicateur ProbLevel
- Nombre d'intervalles
- Entropie
- Entropie mutuelle
- Erreur théorique
- Erreur en apprentissage
- Erreur en validation
- Distance à la loi

Temps de discrétisation

Le temps de discrétisation est mesuré sur un PC Pentium II 500 Mhz, 384 Mo RAM, sous Window/NT 4.0. Seul le temps de discrétisation a été pris en compte. Le chargement des données en mémoire n'a pas été comptabilisé.

ProbLevel

Le ProbLevel est l'indicateur utilisé par la méthode Khiops. Il correspond à la probabilité que la loi de la variable discrétisée et la loi cible soient indépendantes pour la valeur du Khi2 observé (en fait l'opposé du logarithme base 10 pour avoir une valeur positive avec des plages de valeurs facilement interprétables).

Nombre d'intervalles

Le nombre d'intervalles résulte directement de la discrétisation.

Entropie

L'entropie est la mesure de la quantité d'information (nombre de bits) présente dans la variable discrétisée.

Entropie mutuelle

L'entropie mutuelle représente la quantité d'information commune entre la variable discrétisée et la variable cible (qui a ici une entropie de 1).

Erreur théorique

Le prédicteur optimal est celui qui prédit la classe majoritaire en chaque point. Dans notre cas, le prédicteur optimal est basé sur la séparation optimale des deux gaussiennes, c'est à dire sur la médiane des moyennes des deux gaussiennes qui ont même écart type (séparation en 0,5 pour les jeux d'essai). L'erreur théorique correspond au pourcentage de mauvaise prédiction en se basant sur le prédicteur optimal, c'est à dire au rapport de l'aire de l'intersection des deux gaussiennes sur l'aire de la somme des deux gaussiennes.

Erreur en apprentissage

L'erreur en apprentissage est l'erreur mesurée sur la discrétisation si l'on se sert de la discrétisation comme d'un prédicteur. Pour chaque intervalle, on prédit la classe majoritaire mesurée sur l'échantillon. L'erreur en apprentissage est égale à la somme des effectifs des classes minoritaires de chaque intervalle de discrétisation divisée par l'effectif global de l'échantillon. L'erreur en apprentissage résulte donc d'un comptage dont la fiabilité statistique dépend de la taille de l'échantillon.

Erreur en validation

L'erreur en validation de la discrétisation correspond au pourcentage de mauvaise prédiction en se basant sur le prédicteur lié à la discrétisation. Cette erreur en validation peut être calculée de façon exacte car on connaît la loi de distribution exacte de chaque classe. Il n'est pas nécessaire de l'estimer sur un ensemble de validation. Pour cela, on calcule pour chaque intervalle l'aire "erronée" sous la gaussienne de la classe prédite à tort. On fait le cumul de ces aires erronées, divisé par l'aire de la somme des deux gaussiennes. Pour une discrétisation qui se baserait sur exactement deux intervalles $]-\infty; 0,5]$ et $]0,5; +\infty[$ et prédirait la classe 0 sur le premier intervalle et la classe 1 sur le second intervalle, l'erreur en validation coïnciderait avec l'erreur théorique. Cela est logique, car dans ce cas le prédicteur basé sur la discrétisation est égal au prédicteur optimal.

Distance à la loi

On introduit ce dernier critère pour évaluer plus finement que par l'erreur en validation la qualité d'une discrétisation. En effet, si par exemple l'on désire fait du scoring, il faut pouvoir classer les individus par probabilité décroissante d'appartenir à une classe, et donc évaluer cette probabilité plus finement que par une discrétisation à deux intervalles.

Soit une loi de distribution des classes 0 et 1 définie en tout point x de la variable Value par $p_0(x)$ et $p_1(x)$ les probabilités d'appartenir à la classe 0 ou 1, et par la densité de probabilité D(x).

En tout point, $p_0(x) + p_1(x) = 1$ et $\int_{-\infty}^{+\infty} D(x)dx = 1$.

On cherche à comparer cette loi de distribution avec une seconde loi basée sur la même densité de probabilité des individus, mais pour des proportions $p'_0(x)$ et $p'_1(x)$ différentes.

On définit la distance entre les deux distributions de la façon suivante :

$$\text{Distance}\left((p_0, p_1),(p'_0, p'_1)\right) = \frac{1}{2}\int_{-\infty}^{+\infty}\left(|p_0(x) - p'_0(x)| + |p_1(x) - p'_1(x)|\right)D(x)dx$$

Comme il n'y a que deux modalités cibles, on a :

$$\text{Distance}\left((p_0, p_1),(p'_0, p'_1)\right) = \int_{-\infty}^{+\infty}|p_0(x) - p'_0(x)|D(x)dx$$

La distance ainsi définie est donc nulle si et seulement si les deux distributions sont confondues. Elle est bornée par 1 et cette borne est atteinte si par exemple $p'_0(x) = 1-p_0(x)$ et $p_0(x)$ ne prend que des valeurs 0 ou 1.

Graphiquement cette distance peut être vue comme l'aire comprise entre deux distributions (pondérée néanmoins par la densité de la distribution en chaque point). Ceci est illustré ci-dessous pour le cas d'une loi réelle et de sa discrétisation sur la figure suivante.

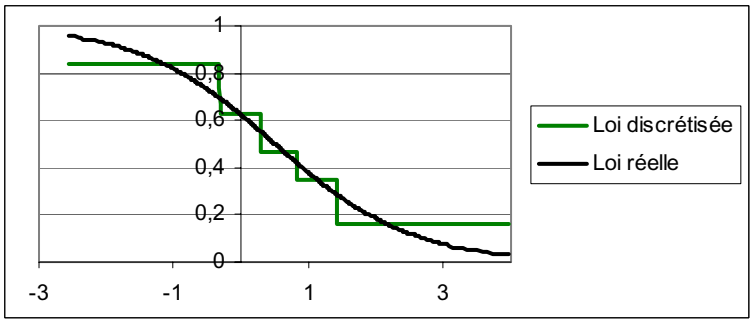

Figure 11 : Visualisation de la distance entre une loi réelle et une loi discrétisée

Dans notre cas, la loi de distribution est basée sur l'utilisation de deux gaussiennes G(x, 0, ET) et G(x, 1, ET).

$p_0(x)$=G(x, 0, ET)/D(x), $p_1(x)$=G(x, 1, ET)/D(x) et D(X) = G(x, 0, ET) et G(x, 1, ET).

La loi de distribution basée sur une discrétisation est la suivante :

- Même densité de probabilité D(x)
- Sur chaque intervalle $]\text{inf}_i; \text{sup}_i]$, on a $p_0(x) = p_{0i}(x)$ calculé par comptage de la proportion dans l'intervalle des individus de la classe 0 en se basant sur l'échantillon ayant servi à la discrétisation

$$\text{Distance}\left(LoiReelle, LoiDiscretisee\right) = \sum_i \int_{\text{inf}_i}^{\text{sup}_i}\left|\frac{G(x,0,ET)}{D(X)} - p_{0i}(x)\right|D(x)dx$$

$$\text{Distance}\left(LoiReelle, LoiDiscretisee\right) = \sum_i \int_{\text{inf}_i}^{\text{sup}_i}\left|G(x,0,ET)\left(1 - p_{0i}(x)\right) - G(x,1,ET)\,p_{0i}(x)\right|dx$$

Pour la discrétisation optimale à deux intervalles, on a :

$\text{Distance}(LoiReelle, LoiDiscretiseeOpt) = 2\int_{-\infty}^{\frac{1}{2}} G(x,0,ET)dx$. La distance entre la loi réelle et la loi discrétisée optimale à deux intervalles est donc égale à deux fois l'erreur théorique. On voit donc que le critère de distance à la loi permet d'évaluer la qualité d'une discrétisation beaucoup plus finement que le critère de l'erreur en validation. On retrouve ici une mesure mettant en relief le compromis recherché par les méthodes de discrétisation. Elles doivent avoir beaucoup d'intervalles pour coller finement à la loi de distribution réelle, mais aussi s'assurer d'effectifs suffisant par intervalle pour garantir une fiabilité statistique.

4.2. Résultats d'expérimentation

ET	Taille	Temps	ProbLevel	ProbLevel ET	Intervalles	Inter... ET	Entropie	Entropie mutuelle	Erreur théorique	Erreur apprentissage	Erreur validation	Distance à la loi
0,1	100	0,00	22,82	0,00	2,00	0,00	0,99	0,99	0,000	0,000	0,000	0,000
0,1	1000	0,00	218,75	0,00	2,00	0,00	1,00	1,00	0,000	0,000	0,000	0,000
0,1	10000	0,07	2173,57	0,00	2,00	0,00	1,00	1,00	0,000	0,000	0,000	0,000
0,1	100000	1,33	21717,30	0,00	2,00	0,00	1,00	1,00	0,000	0,000	0,000	0,000
0,1	1000000	20,05	217150,00	0,00	2,00	0,00	1,00	1,00	0,000	0,000	0,000	0,000
0,2	100	0,00	22,70	0,32	2,00	0,00	0,99	0,98	0,006	0,001	0,009	0,020
0,2	1000	0,00	215,01	1,66	2,00	0,00	1,00	0,96	0,006	0,004	0,007	0,019
0,2	10000	0,08	2124,58	8,19	2,04	0,20	1,01	0,95	0,006	0,006	0,007	0,019
0,2	100000	1,47	21246,00	30,64	3,82	0,43	1,34	0,97	0,006	0,006	0,007	0,011
0,2	1000000	20,69	213237,00	105,27	5,99	0,84	1,26	0,98	0,006	0,006	0,006	0,003
0,25	100	0,00	21,69	0,96	2,00	0,00	0,99	0,91	0,023	0,013	0,029	0,068
0,25	1000	0,01	202,14	3,09	2,09	0,29	1,03	0,87	0,023	0,020	0,025	0,058
0,25	10000	0,09	2006,01	10,84	3,89	0,40	1,57	0,90	0,023	0,023	0,024	0,028
0,25	100000	1,54	20225,20	28,44	7,25	0,80	1,74	0,91	0,023	0,023	0,023	0,007
0,25	1000000	21,94	202580,00	81,75	16,08	1,01	2,20	0,92	0,023	0,022	0,023	0,002
0,3	100	0,00	19,88	1,42	2,02	0,14	1,00	0,80	0,048	0,036	0,055	0,114
0,3	1000	0,01	183,79	4,21	2,96	0,58	1,41	0,79	0,048	0,046	0,051	0,074
0,3	10000	0,11	1854,74	10,27	5,25	0,64	1,85	0,82	0,048	0,048	0,049	0,019
0,3	100000	1,80	18636,60	36,39	11,13	0,89	2,48	0,82	0,048	0,047	0,048	0,006
0,3	1000000	24,38	186571,00	102,28	24,12	1,34	3,25	0,83	0,048	0,047	0,048	0,003
0,4	100	0,00	16,35	1,79	2,26	0,44	1,12	0,63	0,106	0,086	0,118	0,165
0,4	1000	0,01	148,79	5,67	4,14	0,53	1,87	0,62	0,106	0,103	0,110	0,059
0,4	10000	0,16	1495,69	14,99	7,80	0,81	2,63	0,63	0,106	0,106	0,107	0,018
0,4	100000	2,32	15051,50	47,56	16,23	0,96	3,58	0,64	0,106	0,105	0,106	0,007
0,4	1000000	30,46	150817,00	159,21	33,54	1,40	4,54	0,64	0,106	0,105	0,106	0,004
0,5	100	0,00	13,15	1,69	2,56	0,54	1,29	0,50	0,159	0,137	0,174	0,171
0,5	1000	0,02	118,79	5,35	4,44	0,57	2,05	0,48	0,159	0,155	0,164	0,061
0,5	10000	0,21	1185,59	15,25	9,02	0,73	3,02	0,48	0,159	0,158	0,160	0,022
0,5	100000	2,80	11947,10	57,68	18,40	1,19	4,03	0,49	0,159	0,158	0,159	0,009
0,5	1000000	36,02	119740,00	185,48	38,09	1,46	5,06	0,49	0,159	0,158	0,159	0,004
0,6	100	0,00	10,97	1,96	2,61	0,53	1,31	0,41	0,202	0,176	0,222	0,192
0,6	1000	0,02	94,54	5,61	4,71	0,67	2,15	0,37	0,202	0,198	0,208	0,065
0,6	10000	0,25	947,40	15,26	9,35	0,85	3,15	0,37	0,202	0,201	0,204	0,024
0,6	100000	3,31	9536,58	49,42	19,60	1,21	4,21	0,38	0,202	0,202	0,203	0,010
0,6	1000000	40,44	95622,10	167,28	39,48	1,79	5,22	0,38	0,202	0,202	0,202	0,004
1	100	0,00	5,50	1,71	2,52	0,57	1,22	0,19	0,309	0,282	0,329	0,193
1	1000	0,03	44,05	5,18	4,59	0,75	2,09	0,17	0,309	0,304	0,315	0,073
1	10000	0,35	435,55	16,88	9,00	0,85	3,07	0,16	0,309	0,308	0,310	0,027
1	100000	4,37	4417,63	44,66	18,38	1,28	4,09	0,16	0,309	0,308	0,309	0,011
1	1000000	52,63	44430,90	138,88	38,37	1,57	5,15	0,16	0,309	0,308	0,309	0,005
2	100	0,00	2,66	1,17	2,45	0,73	1,12	0,08	0,401	0,361	0,429	0,196
2	1000	0,03	14,69	3,35	3,90	0,87	1,78	0,05	0,401	0,390	0,411	0,078
2	10000	0,41	125,45	10,24	7,02	0,96	2,63	0,04	0,401	0,400	0,403	0,026
2	100000	4,93	1270,91	32,45	13,92	1,30	3,61	0,04	0,401	0,401	0,402	0,010
2	1000000	58,76	12833,70	106,01	28,12	1,54	4,63	0,04	0,401	0,401	0,401	0,004
10	100	0,00	1,64	0,83	2,67	0,79	1,25	0,05	0,480	0,387	0,495	0,220
10	1000	0,03	2,59	1,15	4,03	1,48	1,69	0,01	0,480	0,452	0,491	0,086
10	10000	0,43	7,50	2,39	6,77	2,28	2,34	0,00	0,480	0,473	0,485	0,033
10	100000	5,31	54,90	7,54	9,74	3,07	2,81	0,00	0,480	0,479	0,481	0,011
10	1000000	61,84	540,85	20,48	14,21	2,66	3,42	0,00	0,480	0,480	0,480	0,004

Tableau 11 : Résultats d'expérimentation de la méthode Khiops pour la séparation de deux gaussiennes

La méthode est très rapide. Le temps de discrétisation est de l'ordre de 2 à 5 s pour 100000 individus et de trente à soixante secondes pour un million d'individus. Les temps mesurés en pratique sont conformes à la complexité théorique de l'algorithme en Nlog(N).

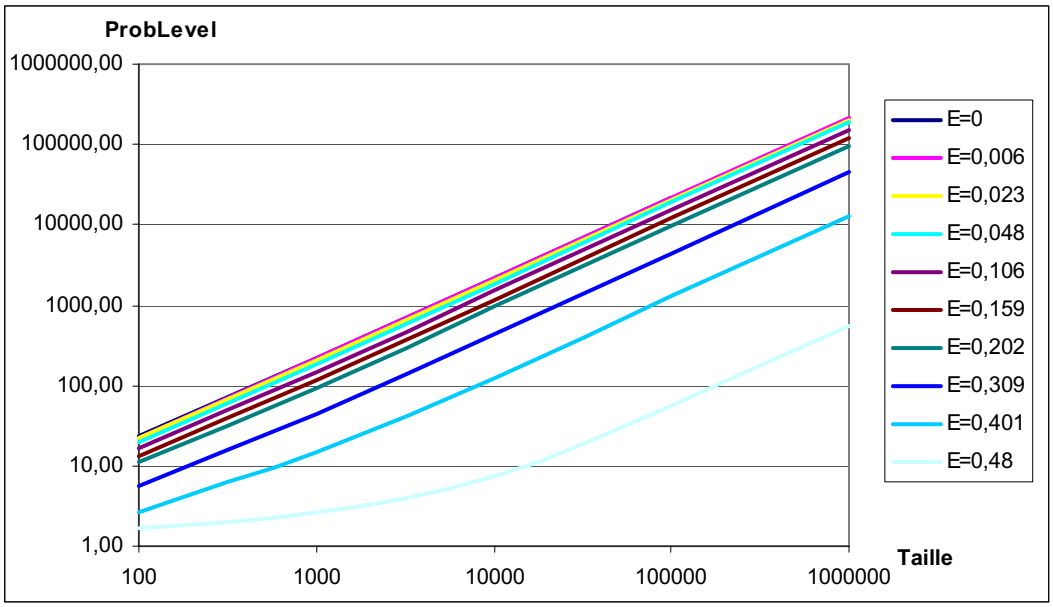

Figure 12 : ProbLevel fonction de la taille de l'échantillon pour des taux d'erreur théorique de 0% à 48%

L'indicateur ProbLevel utilisé par la méthode Khiops est un indicateur stable. Son écart type est très faible devant la valeur de l'indicateur, ce d'autant plus que la taille de l'échantillon est importante. A loi de distribution égale, l'indicateur ProbLevel varie presque linéairement avec la taille de l'échantillon. Son écart type croît plus lentement, environ comme la racine carré de la taille de l'échantillon. A effectif égal, l'indicateur permet d'ordonner parfaitement les jeux d'essai par taux de mélange de gaussiennes décroissant. Sa variation est très corrélée avec les taux de mélange des gaussiennes (mesurée par le taux d'erreur théorique). L'indicateur ProbLevel peut être utilisé comme échelle d'évaluation de la séparabilité des classes pour une variable continue.

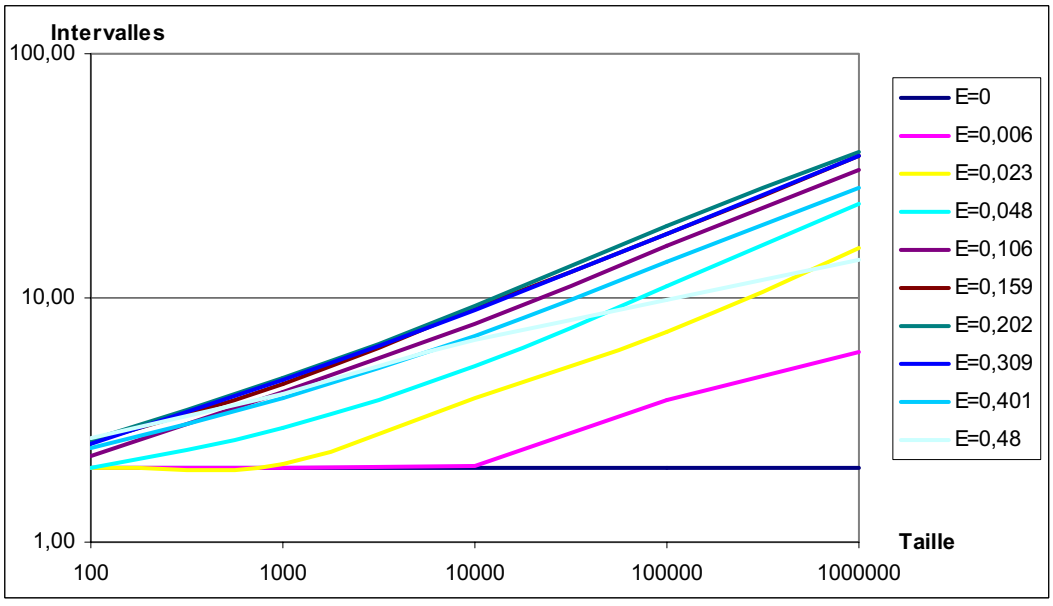

Figure 13 : Nombre d'intervalles fonction de la taille de l'échantillon pour des taux d'erreur théorique de 0% à 48%

Le nombre d'intervalles est également stable (écart type très faible). Il est exactement égal à 2 quand les classes sont parfaitement séparables, croît avec le taux de mélange des classes, puis décroît quand le taux de mélange devient maximal. Il croît avec la taille des échantillons pour permettre de modéliser plus finement la loi réelle, en doublant approximativement quand la taille de l'échantillon est multipliée par

10. Cela permet une discrétisation à la fois plus fine et plus fiable quand la taille de l'échantillon le permet.

L'entropie suit le comportement du nombre d'intervalles. L'entropie mutuelle est fortement corrélée au taux d'erreur optimal, est faiblement dépendante de la taille des échantillons.

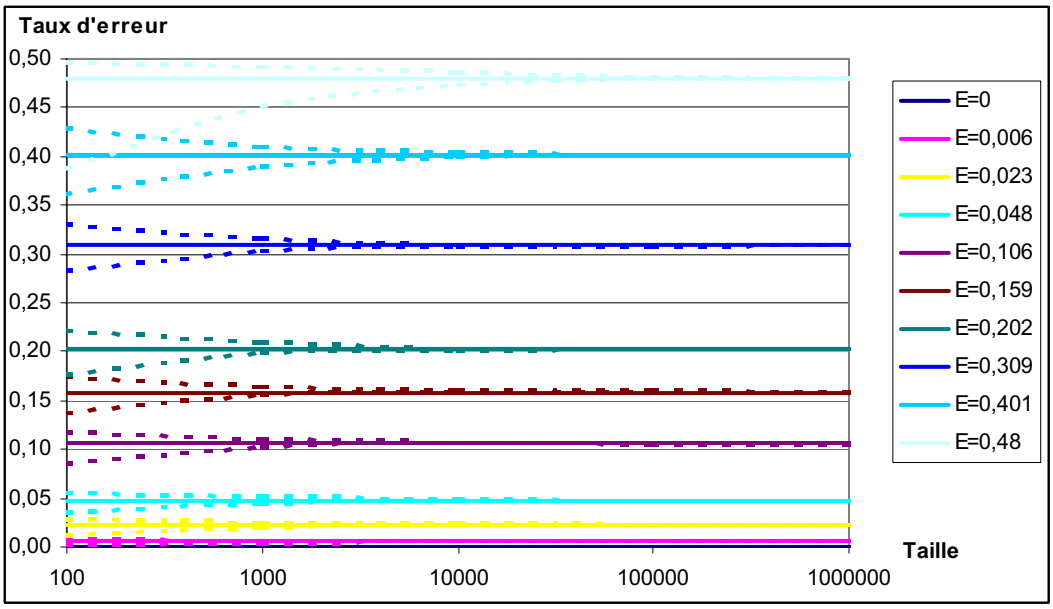

Figure 14 : Erreurs théorique, en apprentissage et en validation fonction de la taille de l'échantillon

L'erreur théorique résulte d'un calcul ne dépendant que de la loi réelle de distribution des mélanges de gaussiennes. On voit ici que les jeux d'essai choisis permettent de balayer les valeurs suivantes de taux d'erreur théorique : 0%, 0,6%, 2,3%, 4,8%, 10,6%, 15,9%, 20,2%, 30,9%, 40,1% et 48%. On a ainsi un éventail très complet des degrés de séparabilité des classes.
La discrétisation peut être vue comme un prédicteur élémentaire basé sur une seule variable, qui donne lieu à une erreur en apprentissage et une erreur en validation. On constate que l'erreur théorique est parfaitement encadrée par l'erreur en apprentissage et l'erreur en validation. L'encadrement est très précis pour tous les jeux d'essai y compris pour le taux d'erreur théorique extrême de 48%. L'écart entre erreur en apprentissage et erreur en validation diminue rapidement avec la taille de l'échantillon. L'écart relatif à l'erreur théorique est raisonnable y compris pour des tailles d'échantillon très petit. Il passe d'environ 25% pour les échantillons de taille 100 à 10% pour des échantillons de taille 1000. Par exemple, le taux d'erreur théorique 10,6% est encadré entre 8,6% et 11,8% pour une taille 100, et entre 10,3% et 11,0% pour une taille 1000.

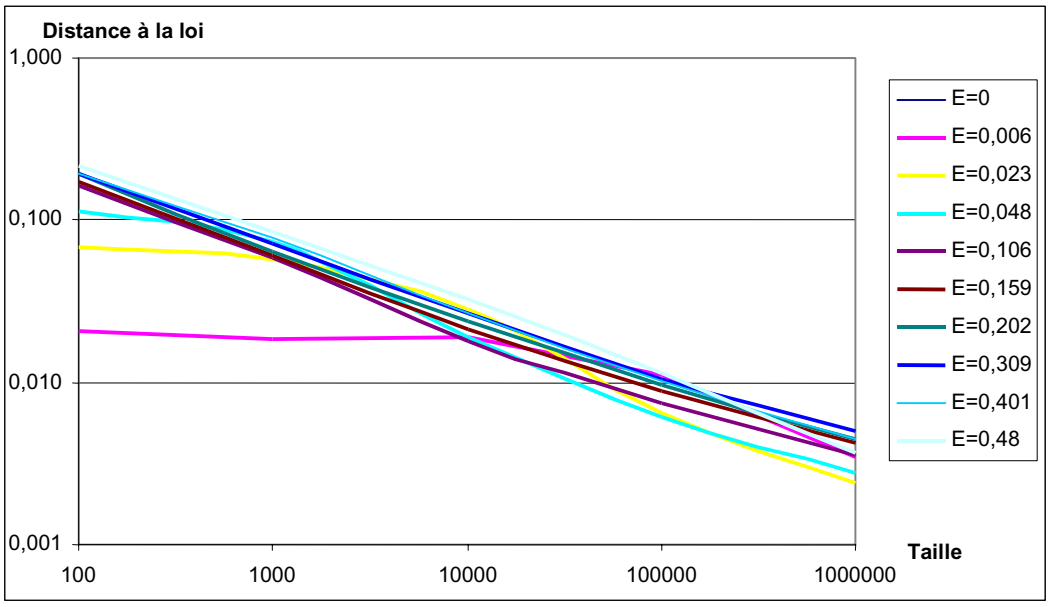

Figure 15 : Distance à la loi réelle de la taille de l'échantillon pour des taux d'erreur théorique de 0% à 48%

La distance entre loi discrétisée et loi réelle est une mesure beaucoup plus fine de la qualité des discrétisations. Dans le cas extrême de classes parfaitement séparables (E=0), la distance à la loi est uniformément nulle. Cette distance croît avec le taux d'erreur théorique. Les distributions de classes peu séparables sont plus difficile à approximer. La distance à la loi décroît rapidement avec la taille de l'échantillon. Cette distance converge toujours vers 0 avec la taille de l'échantillon, quel que soit le taux de mélange des classes. Cette décroissance est approximativement proportionnelle à l'inverse du carré de la taille de l'échantillon.

En ce qui concerne les tests de volumétrie, la méthode Khiops résiste très bien aux volumes de données très importants tant au niveau du temps de calcul qu'au niveau de la qualité des résultats produits, quel que soit le taux de séparabilité des classes.

En résumé, on constate que la méthode de discrétisation Khiops ayant intégré le calcul d'effectif minimum exposé dans cette étude fournit des résultats de grande qualité pour des taux d'erreur théoriques y compris très importants, et ce sans limite de taille pour les échantillons. L'indicateur ProbLevel offre un très bon niveau de stabilité et permet d'évaluer correctement l'importance d'une variable de façon relative à une autre. La fiabilité de la discrétisation, tant pour le taux d'erreur en apprentissage et validation que pour l'estimation correcte de la loi de distribution des classes à prédire, augmente avec la taille des échantillons et tend vers la fiabilité optimale. Le nombre d'intervalles de discrétisation augmente avec la taille de l'échantillon (finesse de discrétisation) et avec le taux de mélange des classes à prédire (complexité de la loi à discrétiser).

4.3. Comparaison avec d'autres méthodes de discrétisation

Afin d'évaluer la méthode, nous avons procédé aux mêmes expérimentations que celles menées dans Zighed et Rakotomalala 2000, dans leur étude sur la discrétisation des attributs continus.

Les auteurs ont utilisé le jeu d'essai Waveform de reconnaissance des ondes proposé par Breiman 1984. Ce jeu d'essai est composé d'un attribut cible comportant 3 classes d'ondes équidistribuées et de 21 attributs continus bruités. Le principe de l'expérimentation est de discrétiser un attribut sur un ensemble d'apprentissage, puis de se servir de la discrétisation pour la prédiction en attribuant à chaque intervalle la classe d'onde majoritaire observée dans l'intervalle sur l'ensemble d'apprentissage. Ce prédicteur est alors évalué sur un ensemble de test en mesurant le taux de reconnaissance. L'expérimentation porte sur 11 échantillons d'apprentissage de 300 points chacun, et un échantillon de test de 5000 points. La discrétisation est donc menée sur 21 attributs dans 11 échantillons, soit 231 fois pour chaque méthode de discrétisation étudiée. La moyenne et l'écart type des taux de reconnaissance pour chaque méthode est alors utilisée comme base de comparaison.

Les méthodes de discrétisation étudiées sont :

- ChiSplit : méthode descendante basée sur le Khi2
- MDLPC : Minimum Description Length Principal Cut (Fayyad 1992)
- Fusbin : mesure d'incertitude utilisée dans la méthode SIPINA (Zighed et Rakotomalala 1996)
- Contrast : critère prenant en compte l'homogénéité des classes et la densité des points (VandeMerckt 1993)
- ChiMerge : méthode ascendante basée sur le Khi2 (Kerber 1991)
- Fusinter : mesure d'incertitude sensible aux effectifs (Zighed et Rakotomalala 1998)
- Fischer : utilisation de l'algorithme (optimal) de Fischer avec le critère Fusinter (Fischer 1958)

Nous avons reproduit ces expérimentations avec la méthode Khiops, et complété ainsi les résultats obtenus par Zighed et Rakotomalala.

Méthode	Moyenne	Ecart type
ChiSplit	0,479323	0,068772
MDLPC	0,481353	0,069429
Fusbin	0,475504	0,069927
Contrast	0,482174	0,069332
ChiMerge	0,442291	0,058811
Fusinter	0,480684	0,069243
Fischer	0,481366	0,069341
Khiops	0,482284	0,069886

Tableau 12 : Moyenne et écart type des taux de reconnaissance estimés pour les différentes méthodes de discrétisation

L'analyse de ces résultats montre que pour le jeu d'essai Waveform, la méthode Khiops se comporte favorablement par rapport aux autres méthodes considérées.

CONCLUSION

La méthode Khiops discrétise une variable continue en minimisant la probabilité d'indépendance entre la loi discrétisée et la loi cible. Cette optimisation est basée sur le critère du Khi2 appliqué de façon globale à l'évaluation de la partition de la variable continue en intervalles, ce qui lui confère des avantages intrinsèques par rapport aux méthodes usuelles apparentées ChiMerge et ChiSplit. Son critère d'arrêt automatique lui confère à la fois une grande facilité d'utilisation et une bonne qualité de la discrétisation obtenue. La complexité algorithmique de la méthode Khiops est la même que pour les méthodes de discrétisation les plus rapides, ce qui a été confirmé lors de tests sur des jeux d'essai de très grande taille.

La méthode Khiops fournit également un indicateur de qualité de la discrétisation obtenue qui est stable vis à vis de l'échantillon discrétisé et constitue un excellent prédicteur du taux de séparabilité des classes à prédire. Cet indicateur permet de comparer l'importance prédictive de variables continues pour une taille d'échantillon donné. Il est particulièrement pertinent pour faire de la sélection de variables.

REFERENCES

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

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

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

W.D. Fischer (1958). « On grouping for maximum of homogeneity ». Jour. Ann. Statis. Assoc. P. 789-798.

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

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

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

T. Vandemerckt (1993). « Decision trees in numerical attributes spaces ». Proceedings oh the 13th IJCAI, Chambery, France.

D.A. Zighed et R. Rakotomalala (1996), « SIPINA-W© for Windows : User's Guide ». Laboratory ERIC – University of Lyon 2.

D.A.Zighed, S. Rabaseda et R. Rakotomalala (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.

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

5. ANNEXE : APPROXIMATION DU DELTAKHI2 POUR LA METHODE KHIOPS

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.

5.1. 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, $\Gamma(n+1) = n!$

$x^{t-1}e^{-x}$

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 :

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

En particulier, on a :

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

5.2. 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.

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

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

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

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

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.

5.3. Calcul du logarithme de probabilité du Khi2

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

5.3.1. 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 : $erfc(x) \approx te^{\left(-x^2+\sum_{i=1}^{9} a_i t^i\right)}$ 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.

5.3.2. Calcul de ln(Q(x,2))

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

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

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

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

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

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

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

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.

5.4. Calcul du DeltaKhi2

5.4.1. 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.

5.4.1.1. 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.

5.4.1.2. 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.

5.4.1.3. 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 $DK(x,n,1) = f_{x,n}(1/n) \sim 1/2\, f_{x,n}(2/n) = 1/2\, DK(x,n,2)$

5.4.2. Calcul de DeltaKhi2 pour un écart de 2 degrés de liberté

5.4.2.1. Bornes de DeltaKhi2

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

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

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

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

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

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.

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

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

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

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

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

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.

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

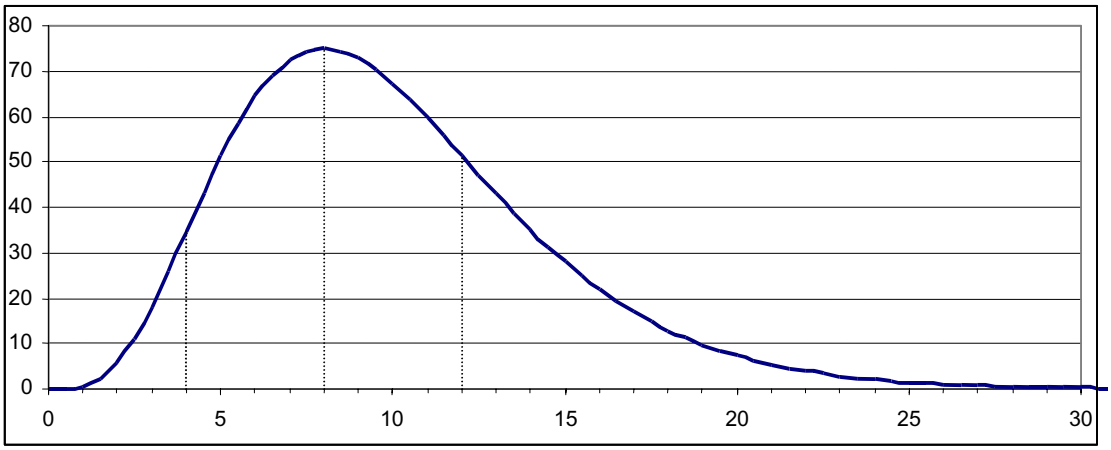

Figure 16 : Courbe f(t) pour n=10

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 $DK(n,n,2) \le \frac{2n}{n-2}$.

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

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

5.4.2.3. 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 :

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

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 $A(x,dx,n)$

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

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

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 $A(x, dx, n) = x/n$

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

5.4.3. 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.

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

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).

5.5. Evaluation numérique

5.5.1. 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.

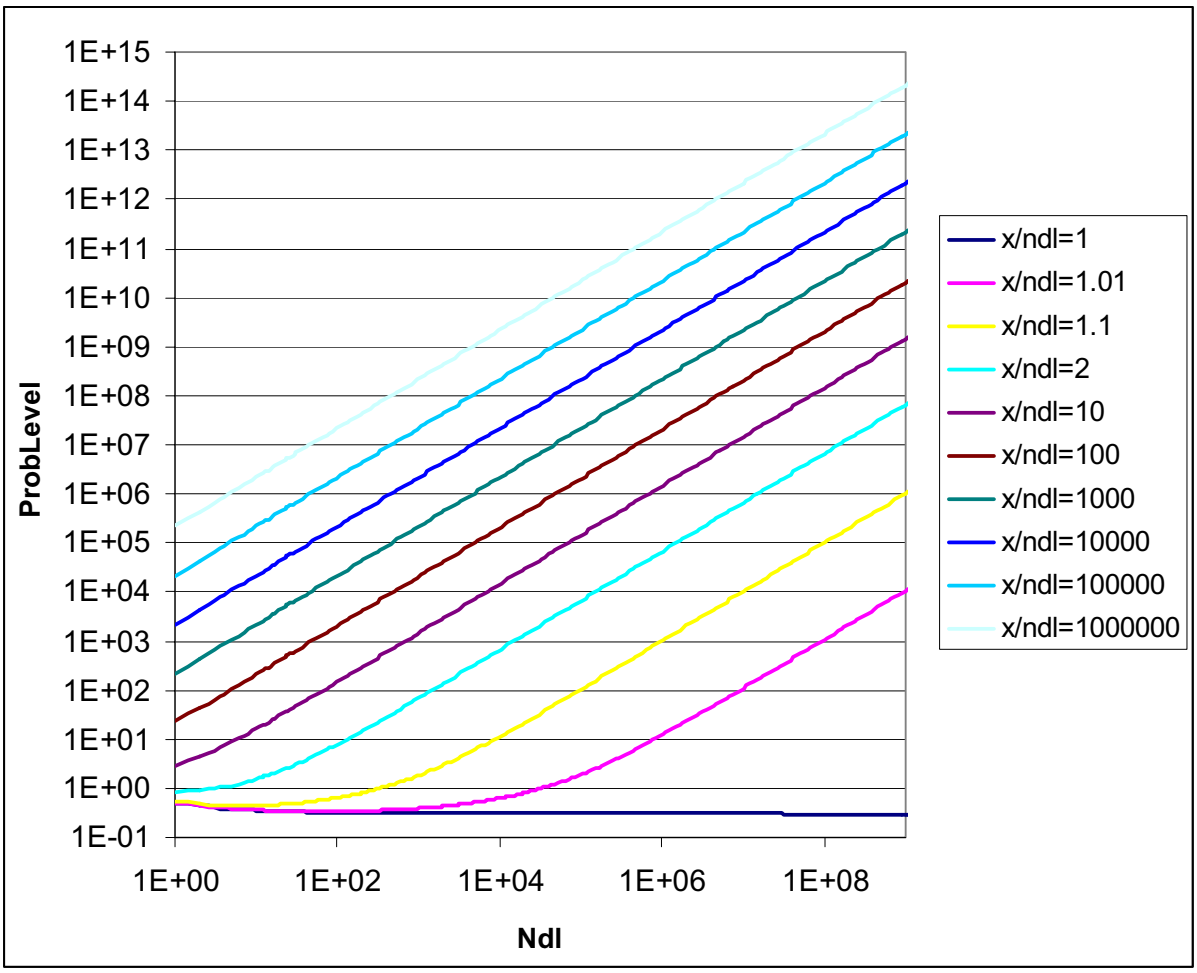

Figure 17 : Logarithme de la probabilité du Khi2 fonction du nombre de degrés de liberté, pour différents ratios x/ndl

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.

5.5.2. 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 $dx = dx_2/2$

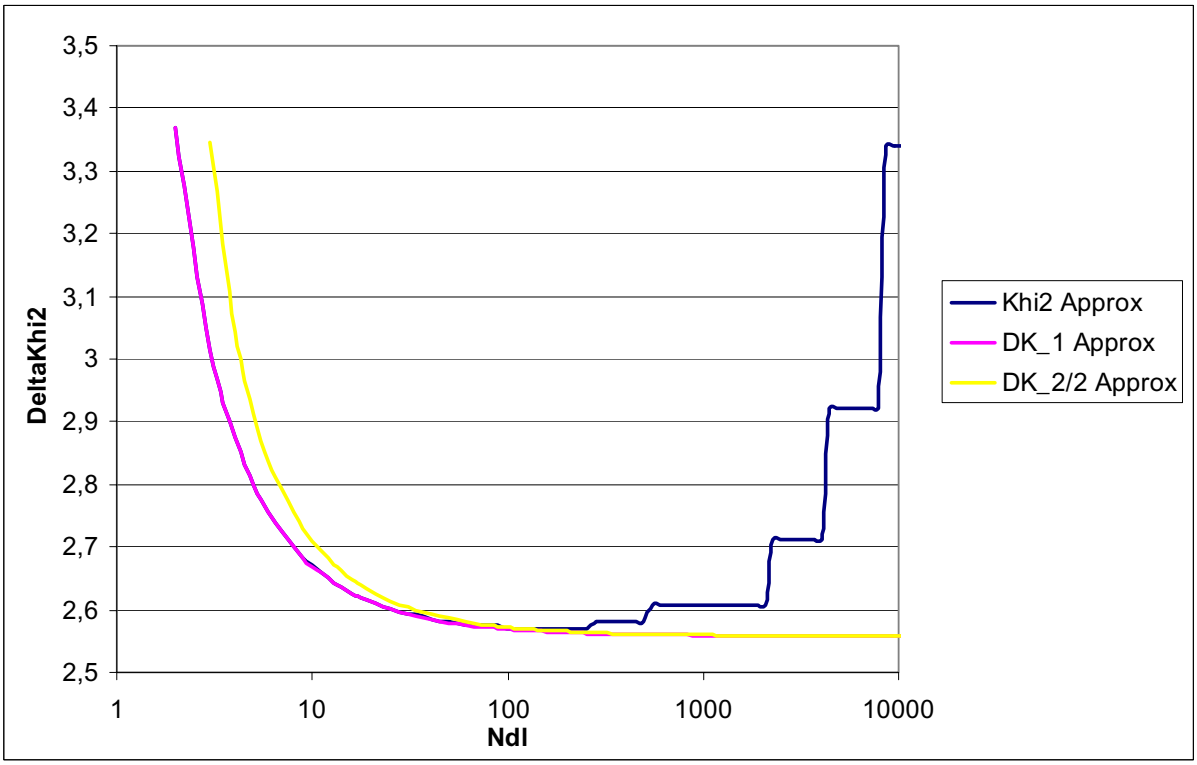

Figure 18 : Comparaison de trois méthodes d'approximation de DeltaKhi2 pour x/ndl=10

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.

5.5.3. 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.

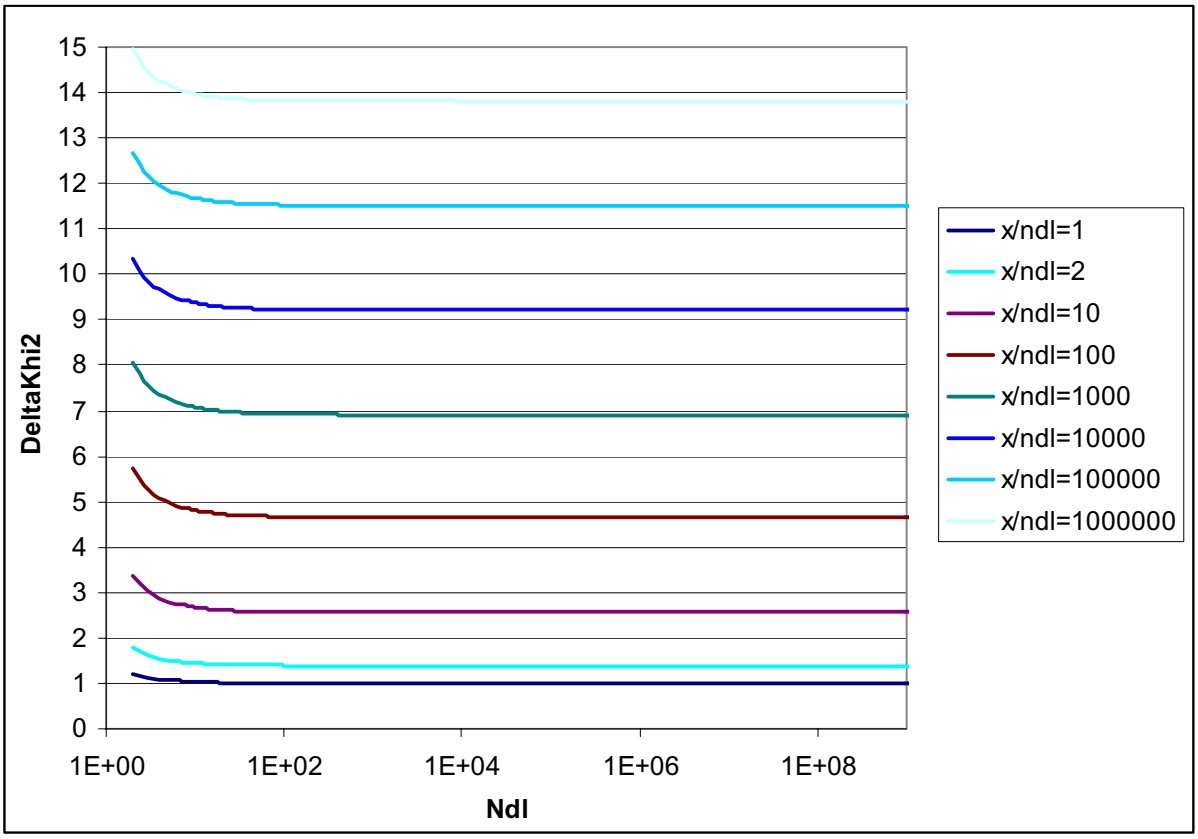

Figure 19 : DeltaKhi2 fonction du nombre de degrés de liberté, pour différents ratios x/ndl

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.

5.6. 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 $$DeltaKhi2 = -\frac{nn'}{n+n'}\frac{(a-b)^2}{p(1-p)}$$

On choisit p=0,5 et on va calculer dans un tableau 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.

	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

Tableau 12 : Valeurs de DeltaKhi2 pour p=0,5 et pour des effectifs observés de 0, 1, 2, 10 et 11

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 $|a-b| \leq \frac{\sqrt{D/2}}{\sqrt{n}}$. 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 $|k-k'| \leq \sqrt{D/2}\sqrt{n}$.

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.

Références

Milton Abramowitz and Irene Stegun (1970), « Handbook of Mathematical Functions », Dover Publications, Ninth Printing.

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