Khiops: une méthode statistique de discrétisation

Marc Boullé

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

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. 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.*

ABSTRACT. *In supervised machine learning, some algorithms are restricted to discrete data. These algorithms thus need to discretize continuous attributes. Many discretization methods have already been studied in the past. They are based on statistical, informational or other specialized criteria. In this paper, we propose a new discretization method Khiops, based on chi-square statistic. Compared with related methods ChiMerge and ChiSplit, this method optimizes the chi-square criterion in a global manner on the entire discretization domain and does not require any stopping criterion. A theoretical study followed with experiments demonstrates the robustness and the predictive accuracy of the method.*

MOTS-CLÉS : *Data Mining, Apprentissage, Discretisation, Analyse de données.*

KEY WORDS: *Data Mining, Machine Learning, Discretization, Data Analysis.*

1. Introduction

La discrétisation des attributs numériques est un sujet largement traité dans la bibliographie (Zighed *et al*, 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 *et al*, 1984) utilise l'indice de Gini (une mesure d'impureté des intervalles), CHAID (Kas, 1980) s'appuie sur une méthode de type ChiMerge (Kerber, 1991), SIPINA (Zighed *et al*, 1996) utilise le critère Fusinter (Zighed *et al*, 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 sensibles aux effectifs et à la distribution de la variable à prédire. Le critère MDL (Minimum Description Length) (Fayyad *et al*, 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 paramétrage par l'utilisateur. En dépit de cette approche globale, l'algorithme associé à la méthode Khiops est de complexité super-linéaire, identique à celle des algorithmes les plus rapides.

Le document est organisé de la façon suivante. 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 pour évaluer la méthode.

2. Méthode de discrétisation 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 libertés). 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 loi discrétisée et loi cible
– Optimisation de la discrétisation: répéter :
  - Evaluer toutes les fusions possibles d'intervalles adjacents
    . Calcul du Khi2 associé à la 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)

L'algorithme Khiops dans son implémentation naïve nécessite N^2 étapes, ou N est le nombre d'individus à discrétiser. En mémorisant les calculs de fusion d'intervalles et utilisant un arbre binaire de recherche pour gérer la liste triée des fusions possibles, on montre que l'on peut se ramener à $N.log(N)$ étapes, ce qui est identique à la complexité de la version optimisée de l'algorithme ChiMerge.

La méthode repose sur l'évaluation de la probabilité du Khi2 pour des valeurs de Khi2 et du nombre de degrés de libertés potentiellement très élevées, ce qui pose problèmes avec les méthodes numériques habituelles. Ces problèmes ont été étudiés en détail et résolus par l'élaboration de méthodes de calcul dédiées (Boullé, 2001).

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.

Dans le cadre de la discrétisation, on procède à des regroupements de valeurs adjacentes en espérant approximer les proportions des modalités cibles à partir des régularités observées dans l'échantillon. Ces régularités proviennent en fait non seulement de la loi de distribution, mais également du hasard lié à l'échantillon. Afin de ne pas se baser à tort sur des régularités qui proviendraient uniquement du hasard, c'est à dire de "sur-apprendre" l'échantillon, une solution est d'augmenter la valeur de l'effectif minimum par intervalle, afin de lisser les effets du hasard. On prendra pour valeur de l'effectif minimum par intervalle la racine carrée de la taille de l'échantillon. Cette contrainte visant à prévenir le sur-apprentissage sera ajoutée à la contrainte d'effectif minimum par case du tableau du Khi2.

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 *et al*, 1998). La base Iris est composée de 150 instances. Les instances représentant des fleurs de la famille des Iris sont décrites par 4 attributs descriptif Sepal Length, Sepal Width, Petal Length, Petal Width et un attribut cible Class comportant 3 modalités (Iris setosa, Iris versicolor, Iris virginica). On va discrétiser l'attribut Sepal Width, qui étant le moins corrélé avec l'attribut cible est le plus intéressant pour illustrer la méthode. Le tableau de contingence associé aux valeurs de l'attribut Sepal Width est le suivant :

Sepal Width	Iris Versicolor	Iris Virginica	Iris setosa	Total
2,0	1	0	0	1
2,2	2	1	0	3
2,3	3	0	1	4
2,4	3	0	0	3
2,5	4	4	0	8
2,6	3	2	0	5
2,7	5	4	0	9
2,8	6	8	0	14
2,9	7	2	1	10
3,0	8	12	6	26
3,1	3	4	5	12
3,2	3	5	5	13
3,3	1	3	2	6
3,4	1	2	9	12
3,5	0	0	6	6
3,6	0	1	2	3
3,7	0	0	3	3
3,8	0	2	4	6
3,9	0	0	2	2
4,0	0	0	1	1
4,1	0	0	1	1
4,2	0	0	1	1
4,4	0	0	1	1
Total	50	50	50	150

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 1. *Evaluation des fusions pour l'attribut Sepal Width de la base Iris*

Lors de l'initialisation, on constitue les 23 intervalles élémentaires]- ∞; 2,1],]2,1; 2,25] …]4,15; 4,3],]4,3; +∞[, comme indiqué dans le tableau 1. La valeur du Khi2 associée est de 88,36. En prenant la loi du Khi2 à 44 degrés de libertés 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 libertés 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.

Sepal Width	Iris versicolor	Iris virginica	Iris setosa	Total
2,0	1	0	0	1
2,2	2	1	0	3
2,3	3	0	1	4
2,4	3	0	0	3
2,5	4	4	0	8
2,6	3	2	0	5
2,7	5	4	0	9
2,8	6	8	0	14
2,9	7	2	1	10
3,0	8	12	6	26
3,1	3	4	5	12
3,2	3	5	5	13
3,3	1	3	2	6
3,4	1	2	9	12
3,5	0	0	6	6
3,6	0	1	2	3
3,7	0	0	3	3
3,8	0	2	4	6
3,9	0	0	2	2
4,0	0	0	1	1
4,1	0	0	1	1
4,2	0	0	1	1
4,4	0	0	1	1
Total	50	50	50	150

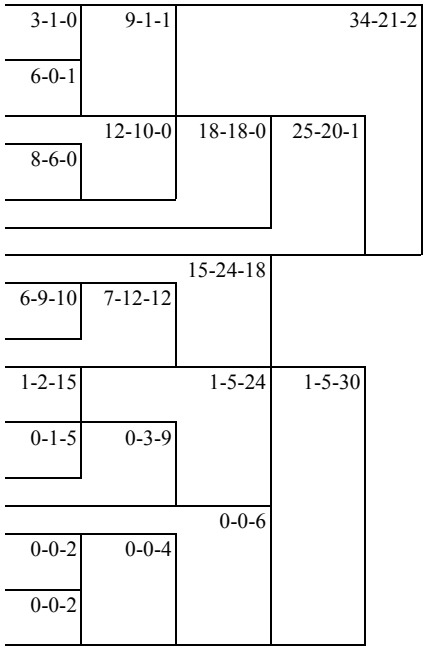

Tableau 2. *Fusions successives jusqu'à une discrétisation en trois intervalles*

Le tableau 2 illustre la liste des étapes successives 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.

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

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 3. *Table de contingence pour l'attribut Sepal Width discrétisé*

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 libertés). 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 libertés). 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.

3. Comparaison théorique avec les méthodes basées sur le Khi2

3.1 *Fusion de deux lignes de Khi2 pour la méthode Khiops*

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

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

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

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

Le Khi2 ligne de la fusion est $Khi2l'' = (n+n')\left(\sum_j \frac{\left((a_j n + b_j n')/(n+n')\right)^2}{p_j} - 1\right)$ [1]

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

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

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

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

3.2 *Comparaison avec ChiMerge*

Pour la méthode ChiMerge, on considère le tableau 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 $q_j = (a_j n + b_j n')/(n+n')$. Pour évaluer l'intérêt de la fusion des deux lignes, on calcule le Khi2 de cette table locale du Khi2.

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

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

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

Pour Khiops, on s'arrête si:

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

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

Proba(SommeKhi2l, J-1) > ProbaSeuil

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

Le tableau 4 illustre la difficulté de choisir un seuil de Khi2 pour ChiMerge.

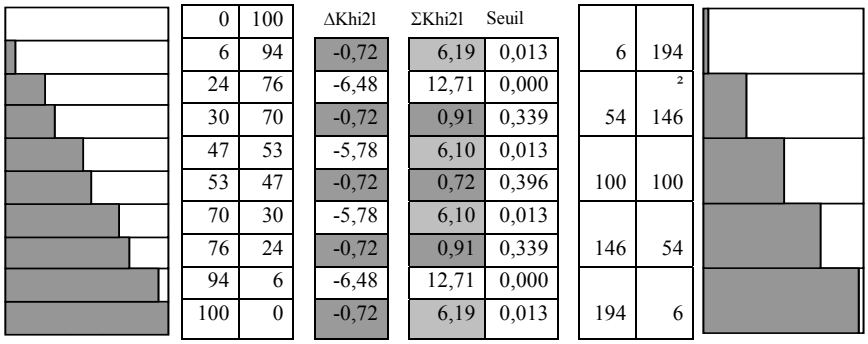

table initiale		Khiops	ChiMerge		table finale	
0	100	ΔKhi2l	ΣKhi2l	Seuil		
6	94	-0,72	6,19	0,013	6	194
24	76	-6,48	12,71	0,000		[2]
30	70	-0,72	0,91	0,339	54	146
47	53	-5,78	6,10	0,013		
53	47	-0,72	0,72	0,396	100	100
70	30	-5,78	6,10	0,013		
76	24	-0,72	0,91	0,339	146	54
94	6	-6,48	12,71	0,000		
100	0	-0,72	6,19	0,013	194	6

Tableau 4. *Choix de la meilleure fusion d'intervalle pour Khiops et ChiMerge*

Dans la table initiale, on a une série de paires d'intervalles ayant des effectifs voisins, et il paraît naturel lors d'une discrétisation de fusionner chaque paire d'intervalles, ce qui correspond au résultat obtenu dans la table finale.

On a ici un Khi2 total pour la table globale de 449,2 égale à environ 50 fois le nombre de degrés de libertés. En se référant à l'étude du calcul numérique des DeltaKhi2 dans (Boullé, 2001), 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. L'étude effectuée dans (Boullé, 2001) montre en fait qu'aucune stratégie pour le choix du critère d'arrêt ne permet de garantir un comportement cohérent tout au long du déroulement de l'algorithme ChiMerge.

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

3.3 *Comparaison avec ChiSplit*

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

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

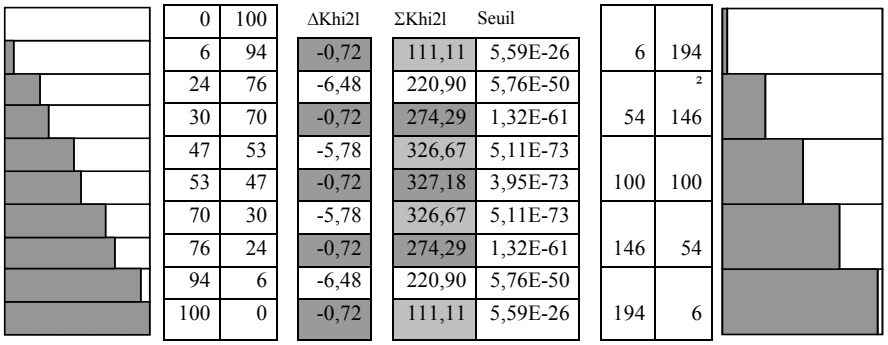

table initiale		Khiops	ChiMerge		Table finale	
		ΔKhi2l	ΣKhi2l	Seuil		
0	100					
6	94	-0,72	111,11	5,59E-26	6	194
24	76	-6,48	220,90	5,76E-50		[2]
30	70	-0,72	274,29	1,32E-61	54	146
47	53	-5,78	326,67	5,11E-73		
53	47	-0,72	327,18	3,95E-73	100	100
70	30	-5,78	326,67	5,11E-73		
76	24	-0,72	274,29	1,32E-61	146	54
94	6	-6,48	220,90	5,76E-50		
100	0	-0,72	111,11	5,59E-26	194	6

Tableau 5. *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 ou plus), on se retrouverait aux limites de la précision numérique des machines (de l'ordre de 10^{-300}), ce qui rendrait impossible le choix d'un seuil. Par ailleurs, la coupure optimale trouvée par ChiSplit est de découper au milieu du tableau du Khi2. En effet, cette coupure donne deux lignes d'effectifs 107-393 et 393-107, qui constitue une excellente coupure de l'ensemble en deux intervalles. Mais de ce fait, la coupure a séparé irrémédiablement les lignes 47-53 et 53-47 qui seraient intuitivement à fusionner. L'approche de l'algorithme ChiSplit qui combine recherche des structures globales et algorithme glouton constitue donc une faiblesse intrinsèque pour l'identification des régularités locales de la variable à discrétiser.

4. Expérimentations

Afin d'évaluer la méthode, nous avons procédé aux mêmes expérimentations que celles menées par (Zighed *et al*, 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 *et al*, 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 chaque 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 *et al*, 1992)

– Fusbin : mesure d'incertitude de la méthode SIPINA (Zighed *et al*, 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 al*, 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	0,4793	0,0687
MDLPC	0,4814	0,0694
Fusbin	0,4755	0,0699
Contrast	0,4822	0,0693
ChiMerge	0,4423	0,0588
Fusinter	0,4807	0,0692
Fischer	0,4814	0,0693
Khiops	0,4823	0,0699

Tableau 6. *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. Il sera toutefois nécessaire de procéder à des tests sur des données réelles et volumineuses afin de poursuivre la validation expérimentale la méthode.

5. 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. La performance prédictive de la méthode a été confirmée par des expérimentations menées sur un jeu d'essai de référence.

Bibliographie

Blake C.L, Merz, C.J., UCI Repository of machine learning databases Available at http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.

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

Boullé M., « Khiops : discrétisation des attributs numériques pour le Data Mining », Note technique NT/FTR&D/7339, France Telecom R&D, 2001.

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

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

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

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

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

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

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

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

Zighed D.A. et Rakotomalala R., *Graphes d'induction*, HERMES Science Publications, 2000.