france telecom R&D

Note Technique
NT/FTR&D/7417

3 octobre 2001

Analyse, sélection et visualisation des couples d'attributs pour l'apprentissage supervisé

Marc Boullé (DTL/DLI)

Vu, pour accord le directeur de DTL

JM. Pitié

Vu, le chef du département DLI

JF. Cloarec

Date : 3 octobre 2001

Résumé : La méthode de discrétisation Khiops optimise le critère du Khi2 globalement sur l'ensemble du domaine de discrétisation et ne nécessite aucun paramétrage. Cette méthode peut être généralisée dans le cadre de l'apprentissage supervisé pour être appliquée au groupage des attributs symboliques, au groupage 2D et à la discrétisation 2D des attributs numériques. Les applications de ces extensions sont nombreuses : visualisation des données, sélection et construction d'attributs, amélioration des performances en modélisation.

Mots clés : **analyse intelligente donnée ; apprentissage automatique ; discrétisation ; visualisation.**
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 **3**
1. MÉTHODE KHIOPS GÉNÉRALISÉE **4**
1.1. RAPPEL DE LA METHODE DISCRETISATION KHIOPS 4
1.2. GENERALISATION 5
2. GROUPAGE DES ATTRIBUTS SYMBOLIQUES **7**
2.1. GROUPAGE 1D 7
2.2. GROUPAGE 2D 7
3. DISCRETISATION 2D **8**
3.1. PRESENTATION 8
3.2. LES DIAGRAMMES DE VORONOÏ DE ET DELAUNAY 8
3.3. METHODE DE DISCRETISATION 2D 9
3.4. EXEMPLE DE DISCRETISATION 2D 10
3.5. EVALUATION DE TOUTES LES PAIRES D'ATTRIBUTS 12
4. APPLICATION A LA PREPARATION DES DONNEES **14**
4.1. PRESENTATION 14
4.2. ETAPE CLASSIQUE DE STATISTIQUES DESCRIPTIVES 14
4.3. APPORTS DE LA METHODE KHIOPS 14
4.3.1. Phase descriptive 14
4.3.2. Sélection des variables 15
4.3.3. Construction de variables 15
4.4. EXEMPLE 15
CONCLUSION **20**
REFERENCES **21**

INTRODUCTION

Les techniques informatiques permettant de réaliser de la fouille de données sont nombreuses et largement répandues : arbres de décision, réseau de neurones, réseaux bayésiens, analyse discriminante… Par contre, les techniques de préparation des données qui représentent jusqu'à 80% du temps d'un cycle du processus Data Mining et sont déterminantes pour la qualité des résultats n'ont pour l'instant été que peu étudiées ou expérimentées. Ces techniques permettent entre autres de sélectionner et de construire des attributs pertinents pour la prédiction.
L'évaluation de l'importance prédictive des attributs est souvent réalisée par des tests statistiques élémentaires : test de Khi2 pour les attributs symboliques, test de Fischer pour les attributs numériques (Saporta 1990). La méthode Khiops (Boullé 2001) offre une alternative intéressante par discrétisation et évaluation des attributs numériques. Nous proposons ici une généralisation de cette méthode qui permet son application à l'évaluation des attributs symboliques et des couples d'attributs symboliques, ainsi qu'une extension en 2D pour l'évaluation des couples d'attributs numériques. Cette analyse des couples d'attributs permet dès la phase de préparation des données de détecter les interactions entre attributs, et d'améliorer ainsi la sélection des attributs. La technique utilisée est non paramétrique, c'est à dire qu'elle peut s'adapter naturellement à de nombreux types d'interaction (contrairement aux techniques de corrélation par exemple).

Le document est organisé de la façon suivante . La partie 1 présente la généralisation de la méthode Khiops. La partie 2 expose son application au groupage des attributs symboliques. La partie 3 introduit l'extension de la méthode pour l'évaluation des paires d'attributs numériques. La partie 4 développe les applications de la méthode pour la préparation des données.

1. METHODE KHIOPS GENERALISEE

1.1. Rappel de la méthode discrétisation Khiops

Le test du Khi2 est à la fois sensible aux effectifs et aux proportions des modalités cibles. Il s'agit donc d'un critère intéressant a priori pour les méthodes de discrétisation. En se basant sur la valeur de la probabilité d'indépendance associée à la loi du Khi2, 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)

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, l'algorithme peut être optimisé pour atteindre 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

1.2. Généralisation

Dans le cas de la discrétisation supervisée, la méthode Khiops peut être vue comme une méthode de réduction du nombre de lignes du tableau du Khi2. Chaque ligne du tableau représente un groupe d'individus. Chaque fusion est possible entre deux groupes d'individus proches, la proximité étant définie selon l'échelle de la variable à discrétiser.

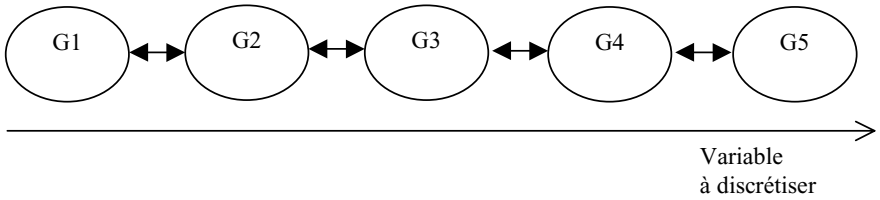

Figure 1 : Discrétisation d'une variable

On peut généraliser ce principe en constituant des groupes d'individus et les fusions possibles entre ces groupes sous forme d'un graphe.

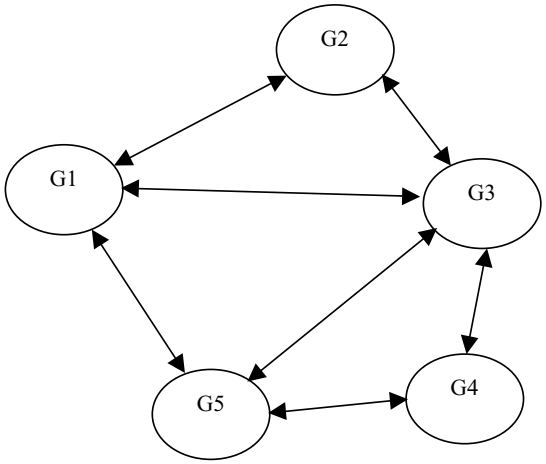

Figure 2 : Graphes des fusions possibles entre groupes d'individus

Chaque groupe représente un sous-ensemble d'individus et correspond à une ligne de tableau du Khi2. Chaque fusion évalue la valeur globale (probabilité d'indépendance) de la partition une fois les deux groupes réunis. La méthode Khiops envisage toutes les fusions possibles et choisit la meilleure. Si cette dernière diminue la probabilité d'indépendance globale entre loi source "partitionnée" et loi cible, la fusion est acceptée. Un nouveau groupe est alors formé et les fusions avec les groupes adjacents sont réactualisées.

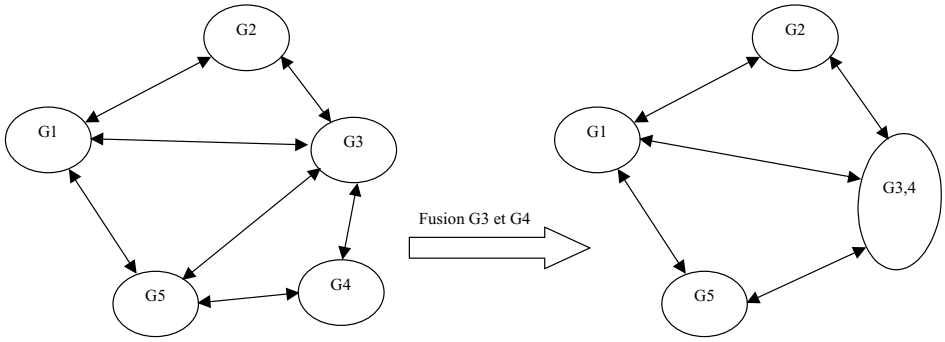

Figure 3 : Fusion de deux groupes lors d'une étape de la méthode Khiops généralisée

Cette extension conduit à l'algorithme suivant.

Algorithme Khiops généralisé

- Initialisation spécifique
  - Constitution des groupes initiaux : G groupes
  - Constitution des fusions initiales : F fusions
- Initialisation
  - Calcul des Khi2 ligne pour chaque groupe et du Khi2 initial : en G
  - Calcul des DeltaKhi2 pour chaque fusion : en F
  - Tri des fusions par valeur de DeltaKhi2 : en Flog(F)
  - Calcul de la probabilité d'indépendance entre la loi discrétisée et la loi cible : en 1
- Optimisation de la discrétisation
  - Répéter: G 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 des groupes
      - Calcul du Khi2Ligne pour le nouveau groupe : en 1
      - Calcul des DeltaKhi2 pour les groupes adjacents au nouveau groupe (au pire : G groupes adjacents)
      - Mise à jour de la liste triée des DeltaKhi2 : en Glog(G)
        - ✓ Suppression du DeltaKhi2 du nouveau groupe
        - ✓ Suppression des anciens DeltaKhi2 des groupes adjacents aux deux sous groupes sources du nouveau groupe
        - ✓ Ajout des nouveaux DeltaKhi2 des groupes adjacents au nouveau groupe

Hormis l'initialisation spécifique, l'algorithme Khiops généralisé a une complexité algorithme en $G^2\log(G)$. Si le graphe des fusions possibles est faiblement maillé et si le nombre d'arcs associé à chaque nœud est borné et le reste au cours des étapes de fusions, la complexité de l'algorithme peut diminuer pour être de l'ordre de Flog(F).

2. GROUPAGE DES ATTRIBUTS SYMBOLIQUES

2.1. Groupage 1D

L'application au groupage des attributs symboliques est immédiate. Chaque groupe initial est constitué de tous les individus ayant même valeurs de modalité source. Toutes les fusions de modalités de l'attribut source sont autorisées.
A la fin de l'algorithme, les modalités sources sont partitionnées en groupes selon le critère défini par la méthode Khiops.

Le groupage supervisé des modalités d'un attribut symbolique est un problème classique. L'apport est ici l'utilisation de la mesure globale de la probabilité d'indépendance et son optimisation par la méthode Khiops. Cette mesure a de solides bases théoriques et permet de comparer des partitions de cardinalités différentes.

2.2. Groupage 2D

Il s'agit d'évaluer si une paire d'attributs symboliques est intéressante pour la prédiction.
On commence ici par fabriquer un nouvel attribut symbolique, produit cartésien des deux attributs symboliques sources. Par exemple, un attribut Taille ayant trois modalités "grand", "moyen" et "petit" sera combiné avec un attribut Poids ayant deux modalités "lourd" et "léger" pour former un nouvel attribut Taille_Poids ayant six modalités "grand_lourd", "grand_léger", "moyen_lourd", "moyen_léger", "petit_lourd" et "petit_léger".
Le groupage 2D est alors identique au groupage 1D appliqué à ce nouvel attribut. Chaque groupe initial est constitué des individus ayant mêmes valeurs de modalités pour les deux attributs symboliques. Toutes les fusions sont autorisées.
A la fin de l'algorithme, les paires de modalités sont partitionnées en groupes selon le critère défini par la méthode Khiops.

Le groupage 2D est largement moins traité dans la bibliographie. On entre ici dans le domaine de la construction d'attribut (feature construction) voire de la modélisation. Le groupage 2D proposé ici bénéficie des apports de la méthode Khiops, et permet de construire un nouvel attribut parmi toutes les combinaisons possibles des modalités des attributs sources.

Il est à noter que cette méthode se généralise naturellement à n attributs symboliques.

3. DISCRETISATION 2D

3.1. Présentation

La discrétisation 2D de deux attributs continus est un sujet très peu traité actuellement. L'application de la méthode Khiops généralisée à ce problème repose sur le même principe général basé sur la proximité des individus Deux groupes d'individus peuvent être fusionnés s'ils sont proches. En discrétisation 1D, la proximité est liée à la distance basée sur la variable source à discrétiser. En 2D, cette proximité provient de la distance euclidienne en prenant les deux variables sources à discrétiser comme axes du plan euclidien.
Chaque individu est représenté par ses coordonnées sur le plan des variables. Comme en discrétisation 1D, chaque individu formera un groupe élémentaire dans l'étape d'initialisation. Il faut alors déterminer les fusions possibles initiales, c'est à dire les relations de proximité entre individus permettant d'envisager leur fusion. Ces fusions sont déterminées à l'aide d'un graphe de Delaunay, qui sera présenté par la suite. La méthode Khiops généralisée est alors appliquée de façon classique sur l'ensemble de ces groupes et fusions initiales.
Cette discrétisation 2D est très prometteuse pour la phase exploratoire du cycle Data Mining. Nous en montrerons plusieurs applications particulièrement intéressantes.

3.2. Les diagrammes de Voronoï de et Delaunay

Le diagramme de Voronoï d'un ensemble de points du plan euclidien permet de partitionner ce plan en cellules de Voronoï. Chaque cellule de Voronoï associée à un point initial est la région du plan où l'on se trouve plus proche de ce point que de tous les autres. Par exemple, si l'on a uniquement deux points initiaux, la médiatrice de ces deux points sépare les points en deux demi-plans qui constitue les cellules de Voronoï associées à chaque point.

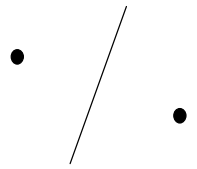

Figure 4 : Diagramme de Voronoï associé à deux points

Dans le cas général de N points initiaux, chaque cellule de Voronoï est un polyèdre convexe dont les segments frontières sont constitués de médiatrices par rapport aux points des cellules de Voronoï adjacentes.

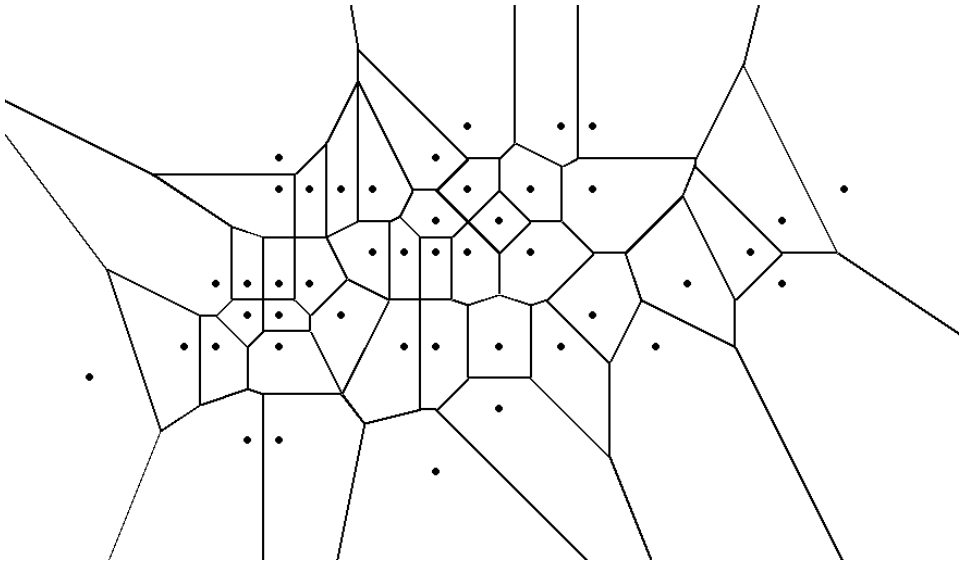

Figure 5 : Diagramme de Voronoï dans le cas général

Le diagramme de Delaunay (ou triangulation de Delaunay) est le diagramme dual du diagramme de Voronoï. Chaque point initial est un nœud de la triangulation de Delaunay. Les arcs relient les nœuds qui sont dans des cellules de Voronoï adjacentes.

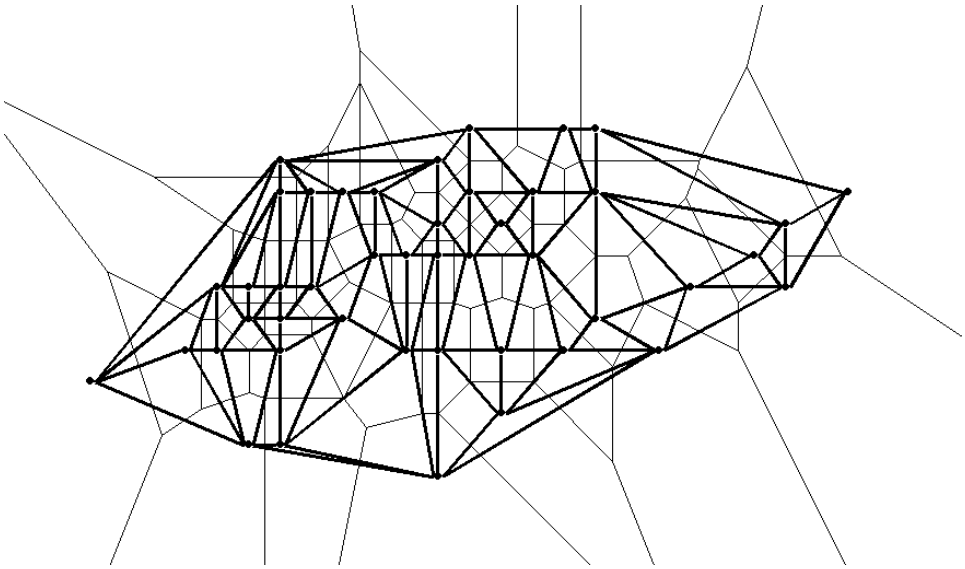

Figure 6 : Diagramme de Delaunay associé au diagramme de Voronoï

Les algorithmes les plus performants utilisés pour le calcul du diagramme de Delaunay de N points ont une complexité algorithmique de Nlog(N). Il y a environ 3N arcs dans une triangulation de Delaunay.

3.3. Méthode de discrétisation 2D

Chaque arc du diagramme de Delaunay représente une relation de proximité entre deux points, ce qui permet de généraliser naturellement la discrétisation en 2D pour deux attributs continus.

Dans l'étape d'initialisation, chaque individu sera considéré comme un point dont les coordonnées sont les valeurs pour les deux attributs à discrétiser en 2D. Chaque point constituera un groupe d'individus élémentaire initial. On calcule alors le diagramme de Delaunay de ces points (complexité en Nlog(N)). Les arcs du diagramme de Delaunay sont utilisés pour initialiser l'ensemble des fusions possibles entre les groupes.

Plus précisément, parmi les arcs du diagramme de Delaunay reliant deux cellules de Voronoï adjacentes, certains sont « directs » et d'autres sont « indirects ». Les arcs directs ne passent que par les deux cellules adjacentes. Le long d'un arc direct, le plus proche voisin est toujours un des deux points des deux cellules adjacentes. Les arcs indirects passent par au moins une troisième cellule de Voronoï. Le long d'un arc indirect, le plus proche voisin peut être un troisième point n'appartenant pas à une des deux cellules adjacentes. Lors d'un prétraitement, les arcs indirects sont éliminés. Seuls les arcs directs, matérialisant complètement une relation de proximités entre points à discrétiser sont pris en compte lors de l'initialisation de la méthode Khiops en 2D.
La méthode Khiops généralisée est alors appliquée sur l'ensemble de ces groupes et fusions initiales. A la fin de l'algorithme, les cellules de Voronoï adjacentes (reliées par un arc de Delaunay) ont été fusionnées selon le critère de la méthode Khiops, et le plan des deux attributs à discrétiser est partitionné en une série de régions connexes. Chaque région regroupe des individus homogènes statistiquement vis à vis de la variable à prédire, et les zones sont par contre différenciées deux à deux. La valeur du critère obtenu à l'issue de la méthode permet de comparer les paires d'attributs continus et de les classer ainsi en fonction de leur valeur prédictive. Le partitionnement des individus en zones connexes sans recouvrement permet une visualisation particulièrement efficace des paires d'attributs les plus pertinentes.

Il est à noter que cette technique se généralise naturellement avec n attributs continus. En effet, le diagramme de Delaunay peut se construire en dimension n.

3.4. Exemple de discrétisation 2D

On va discrétiser l'ensemble des individus de la figure suivante qui sont distribués sur deux classes cibles représentées par des cercles pleins et des diamants.

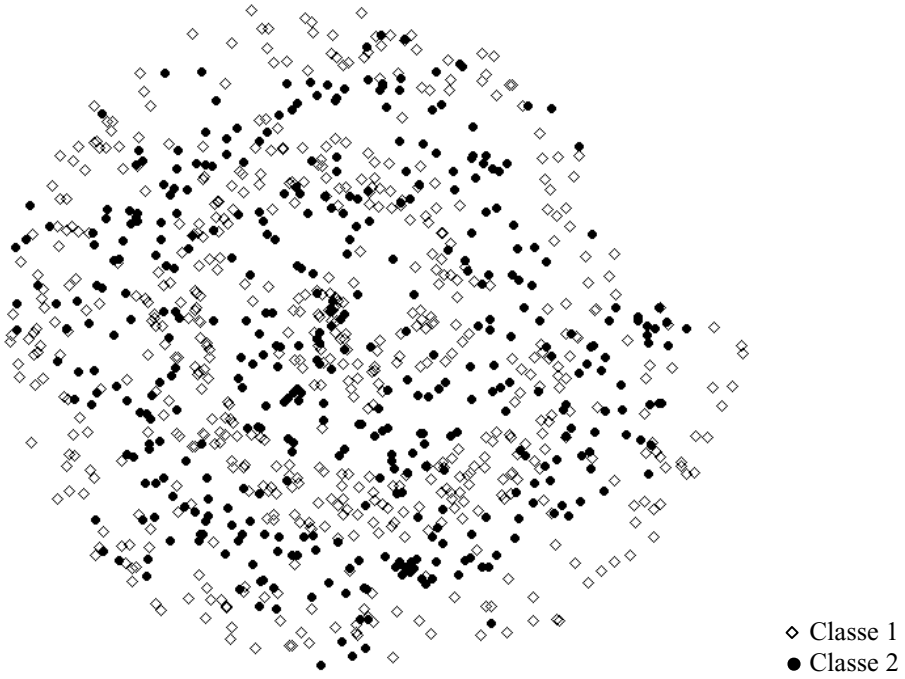

Figure 7 : Individus projetés sur le plan des deux attributs à discrétiser

L'échantillon à discrétiser a une taille totale de 1000 individus, répartis de la façon suivante entre les deux modalités cibles.

	Effectif
Classe 1	589
Classe 2	411

On construit le diagramme de Delaunay des points à discrétiser. On rappelle que l'on ne gardera que les arcs directs de ce diagramme pour initialiser la liste des fusions possibles.

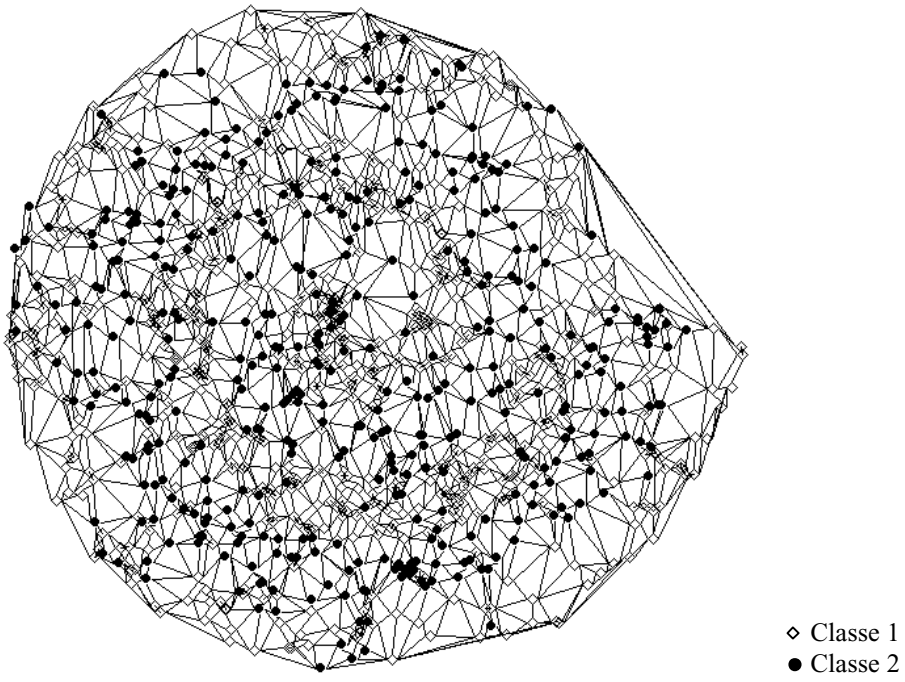

Figure 8 : Diagramme de Delaunay des individus à discrétiser

Après application de la méthode Khiops généralisée, quatre zones ont été identifiées. Chaque point initial est affecté à une des quatre zones pour visualiser la répartition de ces zones.

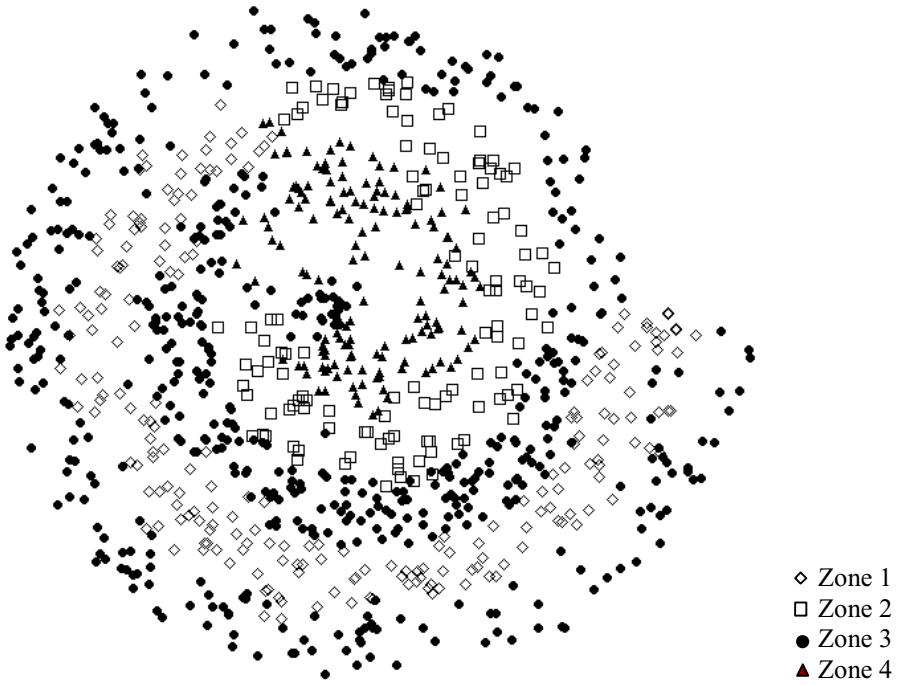

Figure 9 : Zones identifiées par l'algorithme Khiops

Les zones sont homogènes vis à vis du comportement statistique des individus. Le tableau de contingence entre les zones discrétisées et les modalités cibles est le suivant.

	Classe 1	Classe 2	Effectifs
Zone 1	11,8%	88,2%	212
Zone 2	2,5%	97,5%	122
Zone 3	88,7%	11,3%	512
Zone 4	69,5%	30,5%	154

On peut alors visualiser les zones, en leur attribuant des couleurs dépendant des classes majoritaires de chaque zone et en coloriant les cellules de Voronoï des points de chaque zone.

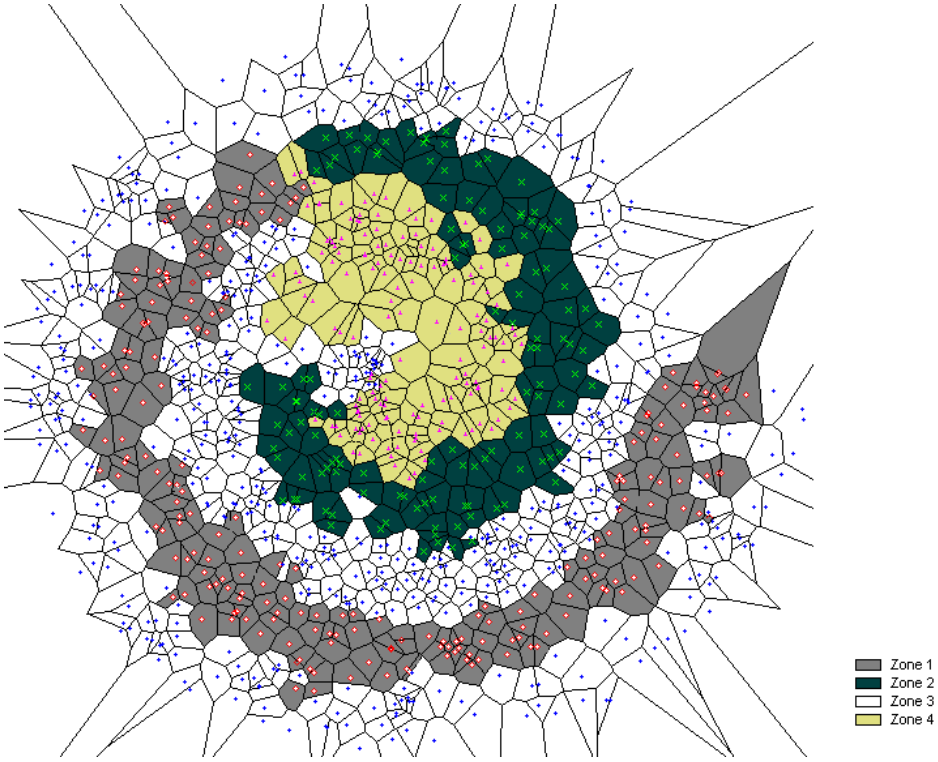

Figure 10 : Visualisation des zones identifiées par l'algorithme Khiops

On observe ainsi très clairement une relation de dépendance en spirale entre les attributs à discrétiser. En fait, le jeu d'essai illustrant cet exemple a été généré artificiellement pour obtenir une dépendance en spirale entre les attributs à discrétiser, en rajoutant 10% de bruit (individus au hasard liés à la mauvaise modalité cible).
La visualisation permet de retrouver cette relation de dépendance, alors qu'elle était moins apparente dans la figure initiale. Le fort taux de bruit rend en effet difficile l'identification visuelle des régularités dans la distribution des points, alors que l'algorithme de discrétisation 2D Khiops parvient à identifier des zones pertinentes pour la prédiction.

La méthode de discrétisation 2D Khiops permet ainsi d'identifier et de visualiser automatiquement des relations de dépendances même complexes entre paires d'attribut continus qui conjointement sont importants pour la prédiction de l'attribut cible.

3.5. Evaluation de toutes les paires d'attributs

Le nombre de paires d'attributs à tester est de l'ordre de la moitié du carré du nombre d'attributs. Quand les attributs sont trop nombreux, il n'est pas possible d'évaluer toutes les paires en temps raisonnable. Une solution à ce problème est de ne rechercher qu'un sous ensemble des paires possibles, et d'utiliser des techniques d'échantillonnage pour procéder à une présélection des paires les plus intéressantes.

On va présenter un algorithme de recherche des K paires d'attributs les plus importantes.
Soit N la taille de la population à évaluer.
Soit N_0 la taille minimale d'échantillon utilisable pour une évaluation (par exemple, on peut prendre N_0=1000).
Soit A le nombre d'attributs à évaluer par paires.
Soit K le nombre de paires à évaluer et classer.

Algorithme d'évaluation des K premières paires d'attributs

- Initialisation
  - Cacul du nombre d'étapes E
    - $E = \log_2(N/N_0)$, c'est à dire $N <= N_0 2^E$
  - Choix d'un échantillon de taille N_0
  - Calcul du nombre max de paires à évaluer à cette étape
    - $k = K*2^E$ ($k=K*N/N_0$)
  - Calcul du nombre d'attributs élémentaires intervenant dans les paires
    - Si k trop petit pour envisager l'évaluation de toutes les paires, utiliser uniquement les attributs les plus importants (de l'ordre de racine de k attributs parmi les A attributs)
- Répéter pour chaque étape
  - Evaluation des k paires d'attributs sélectionnées
  - Préparation de l'étape suivante
    - Choisir un échantillon de taille double
    - Sélectionner la moitié des paires selon l'ordre d'importance

Le nombre d'étapes est de l'ordre de $\log_2(N/1000)$ (si l'on prend N_0=1000). Le nombre total de paires évaluées pour retenir les K meilleures sera de l'ordre de $K*2^E$, ce qui permet une sélection efficace. A chaque étape, on évalue deux fois moins d'attributs sur des échantillons deux fois plus volumineux. Chaque étape a donc une complexité algorithmique approximativement constante, de l'ordre de K fois le temps d'évaluation d'une paire sur la population totale.
En résumé, avec un temps de l'ordre de $K*\log_2(N/1000)$ fois le temps d'évaluation d'une paire d'attributs, l'algorithme présenté permet de sélectionner les K paires les plus importantes de façon statistiquement fiable, en ayant examiné en fait K*N/1000 paires d'attributs.

4. APPLICATION A LA PREPARATION DES DONNEES

4.1. Présentation

Dans les problèmes réels de data mining supervisé, on se trouve fréquemment dans le cas de bases de données comportant des centaines de milliers d'individus, décrits chacun par des centaines d'attributs. Une première étape de statistiques descriptives permet de mieux connaître les attributs décrivant les individus. A l'issue de cette étape, il est nécessaire de sélectionner une partie des attributs, de construire des nouveaux attributs afin d'exploiter au mieux les techniques de modélisation dont les performances se dégradent en général très rapidement avec le nombre d'attributs sélectionnés, le taux de bruit, les attributs redondants…

4.2. Etape classique de statistiques descriptives

Pour chaque attribut symbolique, on calcule la répartition des individus par modalité, le tableau de contingence entre modalités sources et modalités de l'attribut à prédire, ainsi qu'un test du Khi2 permettant d'évaluer si les attributs sont indépendants ou non.
Pour chaque attribut numérique, on calcule des indicateurs statistiques classiques (minimum, maximum, moyenne et écart type) et on effectue un test de Fisher pour évaluer si les différences de moyenne observées pour chaque modalité cible sont dues au hasard ou à une relation de dépendance.
A l'issue de cette étape de statistiques descriptives, on a pour chaque attribut un résumé de l'ensemble des valeurs ainsi qu'un indicateur permettant d'estimer si l'attribut est indépendant avec l'attribut cible ou non.

Pour les attributs numériques, on peut visualiser les histogrammes par modalités cible pour chercher à identifier les relations fortes entre attributs source et cible. Cela n'est exploitable que pour quelques dizaines d'attributs.
On peut également visualiser les paires d'attributs numériques, on affichant sur des diagrammes 2D la distribution des modalités cibles. En pratique, cela n'est possible que dans les cas « simples ». Dans le cas ou l'on se trouve avec plusieurs centaines d'attributs numériques, il y a des dizaines de milliers de paires d'attributs qu'il n'est pas possible de passer en revue même partiellement. De plus, dans le cas de bases assez volumineuses (10000 individus et plus) où certaines modalités sont rares (par exemple, moins de 5% de Churners (clients infidèles) dans une base de clients), il est pratiquement impossible d'identifier visuellement des zones de densité différentes (par exemple, une zone avec 2% de Churners et une zone avec 10% de Churners apparaissent toutes deux noyées parmi les individus non Churners).

4.3. Apports de la méthode Khiops

4.3.1. Phase descriptive

La méthode Khiops permet de procéder au groupage des attributs symboliques et à la discrétisation des attributs numériques automatiquement. A l'issue de cette étape, Khiops produit un indicateur qui permet de classer les attributs par importance prédictive décroissante. De plus, pour chaque variable numérique, on dispose des intervalles de discrétisation qui permettent d'étudier finement l'influence de chaque variable.
Par rapport à la méthode habituelle, la discrétisation automatique permet d'avoir une connaissance beaucoup plus fine des données numériques. De plus, toutes les variables sont classées par importance, ce qui facilite grandement le travail du Data Miner.

Le groupage 2D et la discrétisation 2D constituent un apport significatif. Ces techniques permettent d'évaluer automatiquement l'importance conjointe d'une paire de variables. Les paires de variables symboliques ou continues sont classées par ordre décroissant d'importance, ce qui permet de ne se focaliser que sur les paires les plus importantes. Pour les paires de variables continues, la coloration des zones de discrétisation 2D permet de rendre lisible des zones ayant des différences de comportement subtiles vis à vis des modalités cibles. En évaluant automatiquement l'importance des paires de variables continues et en permettant leur visualisation efficace, la technique de discrétisation 2D permet une véritable exploration des paires d'attributs, y compris dans le cas complexe de bases réelles bruitées, volumineuses, avec rareté de certaines modalités cibles.

4.3.2. Sélection des variables

La méthode Khiops permet de classer tous les attributs symboliques, numériques, et toutes les paires d'attributs symboliques ou numériques par importance décroissante pour la prédiction de l'attribut cible.
Ces classements facilitent la sélection des attributs les plus importants. L'indicateur utilisé dans la méthode Khiops est basé sur une évaluation de la probabilité d'indépendance. Il est utilisable pour des comparaisons d'importance y compris pour des attributs ou paires d'attributs de natures différentes. Cela permet par exemple d'évaluer les redondances en détectant les paires de variables qui conjointement n'apportent pas d'informations par rapport à chaque attribut pris isolément.

4.3.3. Construction de variables

Les groupages et les discrétisations 1D constituent naturellement une façon de construire de nouveaux attributs. Des expérimentations menées sur des bases réelles ont montré l'intérêt de ces nouveaux attributs pour la modélisation. Ce prétraitement des données joue un rôle de « débruitage » et de simplification des attributs qui se traduit en pratique par une amélioration des résultats de modélisation.

Les groupages 2D se prêtent également naturellement à la construction de nouveaux attributs. En ce qui concerne la discrétisation 2D, seul l'apport en préparation des données a pour l'instant été exploité. En revanche, la sélection des meilleures paires d'attributs et leur visualisation efficace permettent au Data Miner d'améliorer sa compréhension du problème traité et constitue une aide en orientant ses essais de constructions de nouveaux attributs et de modélisation.

4.4. Exemple

On va illustrer les apports de la méthode Khiops sur la base Wine provenant des bases d'apprentissage de l'UCI Irvine (Blake 1998).
La base Wine est composée de 178 instances. Les instances correspondent à l'analyse chimique de trois types de vins provenant de la même région. L'analyse a déterminé la quantité de 13 constituants pour chaque type de vin. Les attributs correspondants sont des attributs numériques notés V1, V2… V13 et l'attribut à prédire est un attribut Class pouvant prendre les valeurs 1, 2 ou 3. L'entropie (quantité d'information) de l'attribut Class est 1,56.

La discrétisation Khiops lancée sur l'ensemble des variables V1 à V13 permet de classer ces variables et de déterminer que la variable V7 est de loin la plus importante.

	Discrétisation Khiops			Statistiques				
Attribut	Nombre d'intervalles	ProbLevel	Entropie mutuelle	Nombre de valeurs	Min	Max	Moyenne	Ecart type
V7	3	44,51	0,93	132	0,34	5,08	2,03	1,00
V12	2	29,07	0,60	122	1,27	4	2,61	0,71
V1	2	26,71	0,62	126	11,03	14,83	13,00	0,81
V13	2	26,22	0,54	121	278	1680	746,89	314,02
V10	2	26,05	0,58	132	1,28	13	5,06	2,31
V11	3	25,80	0,58	78	0,48	1,71	0,96	0,23
V6	2	21,08	0,50	97	0,98	3,88	2,30	0,62
V9	3	15,97	0,34	101	0,41	3,58	1,59	0,57
V2	2	14,91	0,28	133	0,74	5,8	2,34	1,11
V4	2	13,90	0,28	63	10,6	30	19,49	3,33
V5	2	11,56	0,26	53	70	162	99,74	14,24
V8	2	10,72	0,22	39	0,13	0,66	0,36	0,12
V3	2	2,77	0,05	79	1,36	3,23	2,37	0,27

Tableau 1 : Discrétisation Khiops des attributs de la base Wine

En observant le tableau de contingence entre la variable V7 et la variable à prédire Class, on observe effectivement une très forte relation de dépendance entre les deux variables.

V7xClass	1	2	3	Total
] -∞ ; 1.4[	0	10	47	57
[1.4 ; 2.31[	1	42	1	44
[2.31 ; +∞[	58	19	0	77
Total	60	73	51	178

Tableau 2 : Tableau de contingence entre la variable V7 et la variable Class

En lançant la discrétisation 2D sur l'ensemble des paires d'attributs et en classant celles ci par indicateur ProbLevel décroissant (indicateur de la méthode Khiops), on peut évaluer l'importance conjointe des paires de variables. Le tableau suivant reproduit ces résultats pour les trente premières paires (parmi 78 paires possibles).

		Discrétisation 2D		
Attribut 1	Attribut 2	Nombre de zones	ProbLevel	Entropie mutuelle
V1	V7	3	70,17	1,43
V6	V10	3	65,63	1,33
V7	V10	3	64,00	1,32
V1	V12	3	63,18	1,26
V1	V11	3	63,15	1,27
V10	V12	3	61,43	1,24
V9	V10	3	61,32	1,24
V7	V11	3	60,90	1,21
V4	V7	3	59,62	1,17
V3	V7	3	59,07	1,19
V7	V9	3	58,87	1,16
V2	V7	3	58,05	1,12
V7	V12	3	57,91	1,14
V2	V12	3	57,80	1,13
V1	V6	3	57,48	1,14
V3	V12	3	56,14	1,08
V1	V10	3	55,75	1,09
V2	V10	3	55,50	1,10
V7	V8	3	52,33	1,05
V6	V11	3	51,62	1,02
V1	V2	3	49,56	0,97
V6	V12	3	49,15	1,00
V6	V7	3	48,46	0,98
V1	V8	3	46,98	0,92
V8	V11	3	46,67	0,92
V1	V9	3	44,79	0,87
V3	V6	3	44,15	0,87
V2	V6	3	43,37	0,85
V2	V11	3	43,04	0,88
V3	V11	3	42,48	0,86
V11	V12	3	38,97	0,86

Tableau 3 : Discrétisation Khiops 2D des paires d'attributs de la base Wine

On peut voir que les 25 premières paires (jusqu'à ProbLevel d'environ 44) sont plus « parlantes » conjointement que les attributs pris individuellement. La première paire V1xV7 se détache nettement des autres. Elle fait intervenir les 1er et 3ème attributs les plus « parlants ». La seconde paire V6xV10 est basée sur les 5ème et 7ème attributs, ce qui est plus remarquable. L'attribut V13 n'intervient dans aucune des 30 premières paires, ce qui le rend à priori moins intéressant que sa 4^{ième} position dans la liste des attributs les plus « parlants ».

En examinant les tableaux de contingence pour les zones de discrétisation des paires V1xV7 et V6xV10, on remarque que les paires apportent nettement plus d'informations que l'attribut V7 seul. La paire V1xV7 arrive à séparer presque parfaitement les trois modalités cibles.

(V1xV7) x Class	1	2	3	Total
Zone 1	0	67	0	67
Zone 2	0	2	48	50
Zone 3	59	2	0	61
Total	59	71	48	178

Tableau 4 : Tableau de contingence entre les zones de discrétisation 2D de V1xV7 et Class

(V6xV10) x Class	1	2	3	Total
Zone 1	0	63	0	63
Zone 2	59	4	0	63
Zone 3	0	4	48	52
Total	59	71	48	178

Tableau 5 : Tableau de contingence entre les zones de discrétisation 2D de V6xV10 et Class

On peut alors procéder à la visualisation des zones de discrétisation 2D. Notons que dans ce cas simple, les zones de discrétisation sont pratiquement pures vis à vis des modalités à prédire.

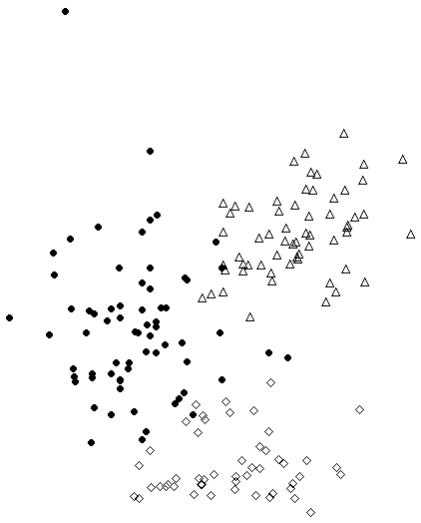

Figure 11 : Zones de discrétisation 2D de V1xV7

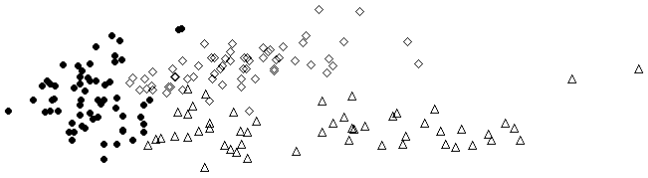

Figure 12 : Zones de discrétisation 2D de V6xV10 (axes inversés)

On observe que les classes sont relativement bien séparables en 2D, alors que sur chacun des axes il y avait effectivement d'importants intervalles avec mélange des modalités cibles.

Afin de vérifier que les meilleures paires sont effectivement classées en tête, on va visualiser la 30^{ième} paire V11xV12 qui fait intervenir les 2^{ième} et 6^{ième} attributs (mieux classés individuellement que pour la paire V6xV10).

(V11xV12) x Class	1	2	3	Total
Zone 1	0	0	41	41
Zone 2	1	33	7	41
Zone 3	58	38	0	96
Total	59	71	48	178

Tableau 6 : Tableau de contingence entre les zones de discrétisation 2D de V11xV12 et Class

Le tableau de contingence des zones de discrétisation montre qu'une des trois zones est constituée d'un mélange des deux modalités cibles 1 et 2.
Les figures suivantes visualisent les zones de discrétisation 2D de V11xV12, et la répartition des modalités cibles dans le plan V11xV12.

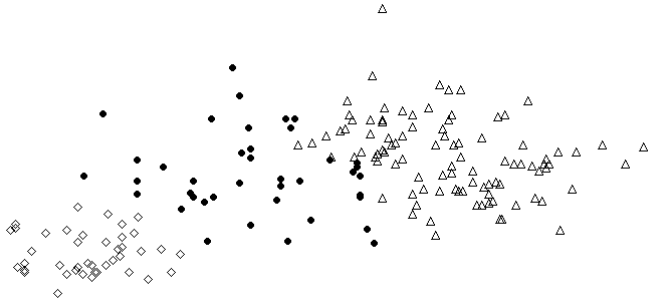

Figure 13 : Zones de discrétisation 2D de V11xV12 (axes inversés)

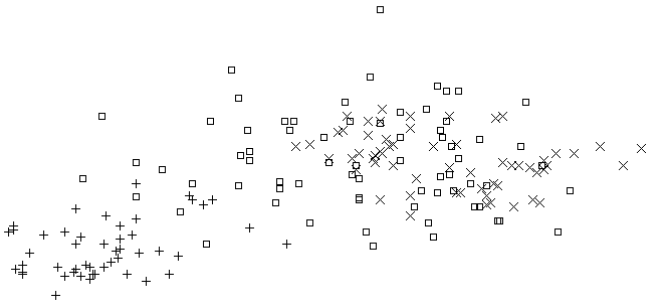

Figure 14 : Répartition des modalités cibles dans le plan V11xV12 (axes inversés)

On constate qu'effectivement, les modalités cibles sont très mélangées, notamment dans la partie droite du plan qui a été regroupée par la discrétisation dans la zone 3. La discrétisation 2D arrive à partitionner le plan V11xV12 en zones homogènes vis à vis des modalités cibles. Son évaluation par l'indicateur ProbLevel reflète la difficulté à trouver un modèle prédictif efficace se basant sur la paire d'attributs V11xV12.

CONCLUSION

La méthode Khiops généralisée permet de proposer une solution au groupage des attributs symboliques, au groupage 2D des paires d'attributs symboliques et à la discrétisation 2D des attributs continus en s'appuyant sur une triangulation de Delaunay préalable des points à discrétiser. Ces techniques peuvent s'étendre naturellement en dimension n. Les applications de la méthode Khiops généralisée concernent essentiellement la phase de préparation des données du Data Mining. Elles permettent d'améliorer rapidement la connaissance des données dans la phase de statistiques descriptives, de classer les attributs et paires d'attributs par ordre d'importance prédictive, de faciliter la sélection des attributs, de visualiser efficacement les données y compris dans les cas complexes, et de construire de nouveaux attributs préparant la phase de modélisation.

RÉFÉRENCES

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.

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

G. Saporta (1990), « Probabilités analyse des données et statistique », Editions TECHNIP.

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