Extraction des Connaissances à partir des Données (ECD)

Data Mining

Par

Djamel Abdelkader ZIGHED & Ricco RAKOTOMALALA

*De document est le « draft » de l'article : Zighed & Rakotomalala, « Extraction des Connaissances à partir des Données (ECD) », in Techniques de l'Ingénieur, **2002**.*

Certaines parties sont clairement obsolètes (extensions, applications : text mining, etc. ; bibliographie) ; d'autres en revanche (fondements méthodologiques, démarche) sont intemporelles.

Remerciements

La rédaction de cet article a été possible grâce au soutien et aux nombreuses contributions des chercheurs

Belkhiter Nadir, Professeur à l'Université Laval à Québec et Professeur invité à l'Université Lumière Lyon 2 pendant l'années 2001-2002.

Hassas Salima, maître de conférences à l'Université Claude Bernard Lyon 1.

Bentayeb Fadila, **Boussaid** Omar, **Darmont** Jérôme, **Rabaséda** Sabine, maîtres de conférences à l'Université Lyon 2 et membres du groupe Bases de Données Décisionnelles du laboratoire ERIC à Lyon 2.

Muhlenbach Fabrice, **Clech** Jérémy chercheurs en thèse au laboratoire ERIC à Lyon 2.

Qu'ils soient tous très chaleureusement remerciés.

Nous tenons également à exprimer nos remerciements à tous les membres du laboratoire ERIC qui ont, par leur encouragement et leur disponibilité, facilité la réalisation de cet article.

1. Introduction générale

Le *data mining*, dans sa forme et compréhension actuelle, à la fois comme champ scientifique et industriel, est apparu au début des années 90. Cette émergence n'est pas le fruit du hasard mais le résultat de la combinaison de nombreux facteurs à la fois technologiques, économiques et même sociopolitiques.

On peut voir le *data mining* comme une nécessité imposée par le besoin des entreprises de valoriser les données qu'elles accumulent dans leurs bases. En effet, le développement des capacités de stockage et les vitesses de transmission des réseaux ont conduit les utilisateurs à accumuler de plus en plus de données. Certains experts estiment que le volume des données double tous les ans. Que doit-on faire avec des données coûteuses à collecter et à conserver ?

Une confusion subsiste encore entre *data mining*, que nous appelons en français « fouille de données », et *knowledge discovery in data bases* (KDD), que nous appelons en français « extraction des connaissances à partir des données » (ECD). Le *data mining* est l'un des maillons de la chaîne de traitement pour la découverte des connaissances à partir des données. Sous forme imagée, nous pourrions dire que l'ECD est un véhicule dont le *data mining* est le moteur.

Le *data mining* est l'art d'extraire des connaissances à partir des données. Les données peuvent être stockées dans des entrepôts (*data warehouse*), dans des bases de données distribuées ou sur Internet. Le *data mining* ne se limite pas au traitement des données structurées sous forme de tables numériques ; il offre des moyens pour aborder les corpus en langage naturel (*text mining*), les images (*image mining*), le son (*sound mining*) ou la vidéo et dans ce cas, on parle alors plus généralement de *multimedia mining*.

L'ECD, par le biais du *data mining*, est alors vue comme une ingénierie pour extraire des connaissances à partir des données.

L'ECD est un processus complexe qui se déroule suivant une série d'opérations. Des étapes de pré-traitement ont lieu avant le *data mining* en tant que tel. Le pré-traitement porte sur l'accès aux données en vue de construire des *datamarts*, des corpus de données spécifiques. Le pré-traitement concerne la mise en forme des données entrées selon leur type (numériques, symboliques, images, textes, sons), ainsi que le nettoyage des données, le traitement des données manquantes, la sélection d'attributs ou la sélection d'instances. Cette première phase est cruciale car du choix des descripteurs et de la connaissance précise de la population va dépendre la mise au point des modèles de prédiction. L'information nécessaire à la construction d'un bon modèle de prévision peut être disponible dans les données mais un choix inapproprié de variables ou d'échantillon d'apprentissage peut faire échouer l'opération.

Le *data mining*, dans sa définition restreinte, opère sur des tables bidimensionnelles, appelées *datamarts*, et fait appel à trois grandes familles de méthodes issues de la statistique, de l'analyse des données, de la reconnaissance de formes ou de l'apprentissage automatique. Il est possible de regrouper les méthodes couramment utilisées ou présentées comme faisant partie de l'arsenal du « *data* miner » en trois catégories :

- les méthodes de description uni, bi et multidimensionnelles : numériques, pour la plupart, elles sont issues de la statistique descriptive et de l'analyse des données, ainsi que des techniques de visualisation graphiques dont certaines font appel à la réalité virtuelle et à des métaphores de représentation assez élaborées ;

- les méthodes de structuration qui regroupent toutes les techniques d'apprentissage non supervisé et de classification automatique provenant

des domaines de la reconnaissance de formes, de la statistique, de l'apprentissage automatique et du connexionnisme ;

- les méthodes explicatives dont le but est de relier un phénomène *à expliquer* à un phénomène *explicatif* : généralement mises en œuvre en vue d'extraire des modèles de classement ou de prédiction, ces méthodes descendent de la statistique, de la reconnaissance de formes, de l'apprentissage automatique et du connexionnisme, voire du domaine des bases de données dans le cas de la recherche de règles d'associations.

En dehors du champ des statisticiens, on assiste à l'émergence d'outils plutôt que de méthodes exploratoires. On peut ainsi citer les algorithmes de recherche de règles d'associations dans les grandes bases de données. Les premiers algorithmes proposés dans ce domaine ont fait sourire des membres de la communauté des statisticiens et des spécialistes de l'induction en raison de la naïveté du matériel méthodologique qui était alors utilisé. Par la suite, ces problèmes ont été ramenés dans un cadre méthodologique plus général, faisant par exemple usage de parcours de treillis de Gallois ou de recherche de décomposition optimale d'une relation binaire par des relations dites maximales.

L'objectif de la mise en œuvre des techniques de *data mining* est d'aboutir à des connaissances opérationnelles. Ces connaissances sont exprimées sous forme de modèles plus ou moins complexes : une série de coefficients pour un modèle de prévision numérique, des règles logiques du type « Si Condition alors Conclusion » ou des instances. Pour que ces modèles acquièrent le statut de connaissances, ils doivent être validés. Il s'agit alors de mettre en œuvre une série d'opérations dites de post-traitement qui visent à évaluer la validité des modèles, à les rendre intelligibles s'ils doivent être utilisés par l'homme ou à les exprimer dans un formalisme approprié pour être utilisé par une machine. Au-delà de la validation statistique, l'intelligibilité des modèles est souvent un critère de leur survie. En effet, un modèle compris par l'usager sera utilisé et par conséquent critiqué et perfectionné. Les utilisateurs n'aiment généralement pas employer de modèles sous forme de « boîtes noires ».

Une question importante, dans le domaine du *data mining*, est de pouvoir répondre à la question : quel outil doit-on employer pour quel problème ? Selon le type de problème, il existe de nombreuses méthodes de *data mining* concurrentes. Un consensus général semble se dégager pour reconnaître qu'aucune méthode ne surpasse les autres car elles ont toutes leurs forces et faiblesses spécifiques. Il semble plus avantageux de faire coopérer des méthodes comme nous le ferions avec une équipe de spécialistes.

Les techniques de *data mining* ont été employées avec beaucoup de succès dans de grands secteurs d'application : la gestion de la relation client (GRC) – ou *customer relationship management* (CRM) –, la gestion des connaissances – *knowledge management* – ou l'indexation de documents. Aucun domaine

d'application n'est *a priori* exclu car dès que nous sommes en présence de données empiriques, le *data mining* peut rendre de nombreux services.

Il existe une large panoplie de logiciels de *data mining* recensés sur Internet. L'un des meilleurs sites de référence est kd.nuggets.com, un excellent portail pour ne pas se perdre dans l'univers du *data mining*.

Le *data mining* est un domaine à la fois scientifique et technologique récent qui a encore de nombreux défis à relever. La communauté des chercheurs dans ce domaine s'intéresse ainsi à des problèmes tels que la recherche de bons espaces de représentation ou à l'agrégation de prédicteurs, etc.

Grâce à Internet, une grande quantité de sites regroupant des logiciels, des données, des expertises, des cours, des communautés d'échanges et de la bibliographie sont à présent accessibles.

La bibliographie dans ce domaine, si elle est encore aujourd'hui limitée en termes d'ouvrages, est cependant très abondante en termes d'articles.

2. Historique

L'expression « *data mining* » est apparue vers le début des années 1960 et avait, à cette époque, un sens péjoratif. En effet, les ordinateurs étaient de plus en plus utilisés pour toutes sortes de calculs qu'il n'était pas envisageable d'effectuer manuellement jusque là. Certains chercheurs ont commencé à traiter sans *a priori* statistique les tableaux de données relatifs à des enquêtes ou des expériences dont ils disposaient. Comme ils constataient que les résultats obtenus, loin d'être aberrants, étaient encourageants, ils furent incités à systématiser cette approche opportuniste. Les statisticiens officiels considéraient toutefois cette démarche comme peu scientifique et utilisèrent alors les termes « *data mining* » ou « data fishing » pour les critiquer.

Cette attitude opportuniste face aux données coïncida en France avec la diffusion dans le grand public de l'analyse de données dont les promoteurs, comme Jean-Paul Benzecri, ont également dû subir dans les premiers temps les critiques venant des membres de la communauté des statisticiens.

Le succès de cette démarche empirique ne s'est malgré tout pas démenti. L'analyse des données s'est développée et son intérêt grandissait en même temps que la taille des bases de données. Vers la fin des années 1980, des chercheurs en base de données, tel que Rakesh Agrawal, ont commencé à travailler sur l'exploitation du contenu des bases de données volumineuses comme par exemple celles des tickets de caisse de grandes surfaces, convaincus de pouvoir valoriser ces masses de données dormantes. Ils utilisèrent l'expression « *database mining* » mais, celle-ci étant déjà déposée par une entreprise (*Database mining workstation*), ce fut « *data mining* » qui s'imposa. En mars 1989, Shapiro Piatetski proposa le terme « *knowledge discovery* » à l'occasion d'un atelier sur la découverte des connaissances dans

les bases de données. Actuellement, les termes *data mining* et *knowledge discovery in data bases* (KDD, ou ECD en français) sont utilisés plus ou moins indifféremment. Nous emploierons par conséquent l'expression « *data mining* », celle-ci étant la plus fréquemment employée dans la littérature.

La communauté de « *data mining* » a initié sa première conférence en 1995 à la suite de nombreux *workshops* sur le KDD entre 1989 et 1994. En 1998, s'est créé, sous les auspices de l'ACM, un chapitre spécial baptisé ACM-SIGKDD, qui réunit la communauté internationale du KDD. La première revue du domaine « *Data mining and knowledge discovery journal* » publiée par « Kluwers » a été lancée en 1997.

3. Valoriser les données

Au-delà de l'origine de la paternité de l'expression « *data mining* », nous allons maintenant nous intéresser à l'émergence de ce champ à la fois technologique et scientifique. L'exploitation des données pour en extraire des connaissances est une préoccupation constante de l'être humain car elle est une condition essentielle de son évolution. L'homme a toujours mémorisé sur des supports différents des informations qui lui ont permis d'inférer des lois. La biologie, la physique, la chimie ou la sociologie, pour ne citer que ces disciplines, font largement usage de l'approche empirique pour découvrir des lois et ou faire ressortir des éléments structurants dans des populations. La statistique est devenue une science dont l'objet est de donner un cadre rigoureux à la démarche empirique. C'est au sein de la statistique et du domaine des bases de données que le *data mining* a puisé ses outils.

Les historiens des sciences, s'ils ne l'ont pas déjà fait, vont certainement nous proposer de nombreuses théories pour mieux situer le *data mining* dans le domaine des sciences.

Dans sa forme actuelle, le *data mining* est né d'un besoin : valoriser les bases de données dont la taille croît de manière exponentielle afin de mieux maîtriser la compétitivité. Par exemple l'exploitation de l'historique des achats des clients permet d'optimiser qualitativement et quantitativement les campagnes de marketing.

4. Les facteurs d'émergence du *data mining*

4.1 Le volume des bases de données

Si nous regardons la jeune histoire de l'informatique, nous pouvons identifier au moins 4 phases :

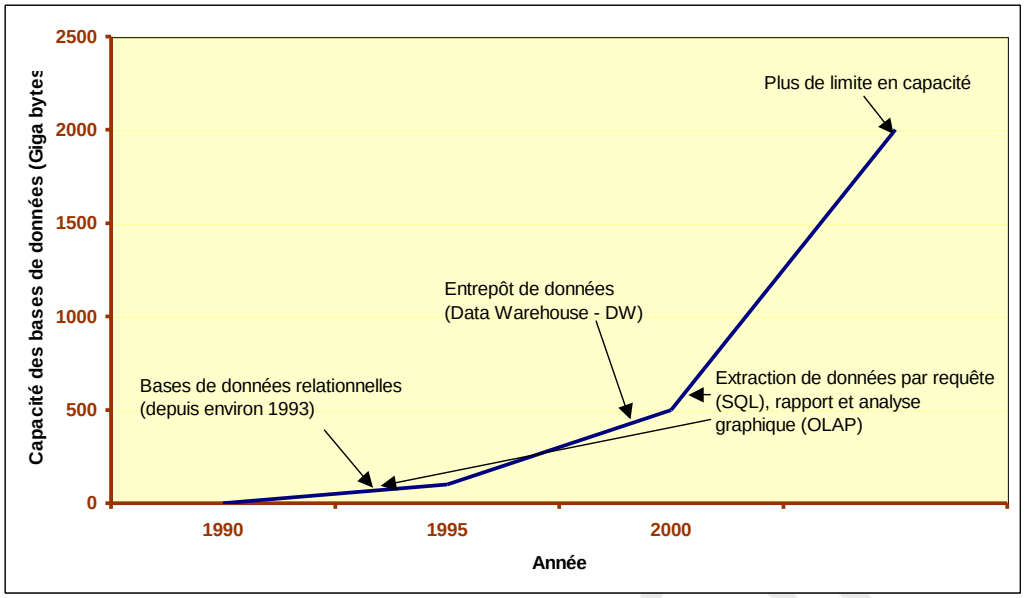
- **Avant 1970**, l'informatique était massivement focalisée sur la recherche des moyens méthodologiques et technologiques pour automatiser de manière

efficiente l'acquisition, la conservation et la restitution des grandes masses de données potentiellement disponibles dans les systèmes d'informations des organisations. Les systèmes de gestion des fichiers (SGF) utilisés dans ce but à cette époque ne fournissaient que de faibles garanties quant à la cohérence sémantique, la non-redondance ou la validité des informations qui se trouvaient sur les ordinateurs. Les programmeurs passaient alors une grande part de leur temps à modifier des programmes parce qu'une donnée nouvelle était introduite dans le système d'information de l'entreprise. Par exemple, un nouveau prélèvement sur les salaires des employés entraînait une cascade de modifications dans la chaîne de traitement des salaires. Ces faiblesses engendraient des coûts de mise en place et de maintenance des systèmes d'information automatisés élevés. L'enjeu était de taille et tous les efforts se consacraient à résoudre le problème du nécessaire passage à une plus grande maîtrise des systèmes d'information.

- **Les années 1970-1980**. Les travaux de Codd sur le modèle relationnel en bases de données ont apporté une réponse dont la pertinence ne s'est encore jamais démentie. Toutes les entreprises qui, jusque là, pataugeaient dans la glaise des systèmes de gestion de fichiers pour assurer la production d'information fiable et cohérente, ont vu apparaître l'outil de productivité tant attendu : les premiers systèmes de gestion de bases de données (SGBD) modernes fournis par des constructeurs comme IBM avec, en prime, les langages de 4ème génération tels que SQL.

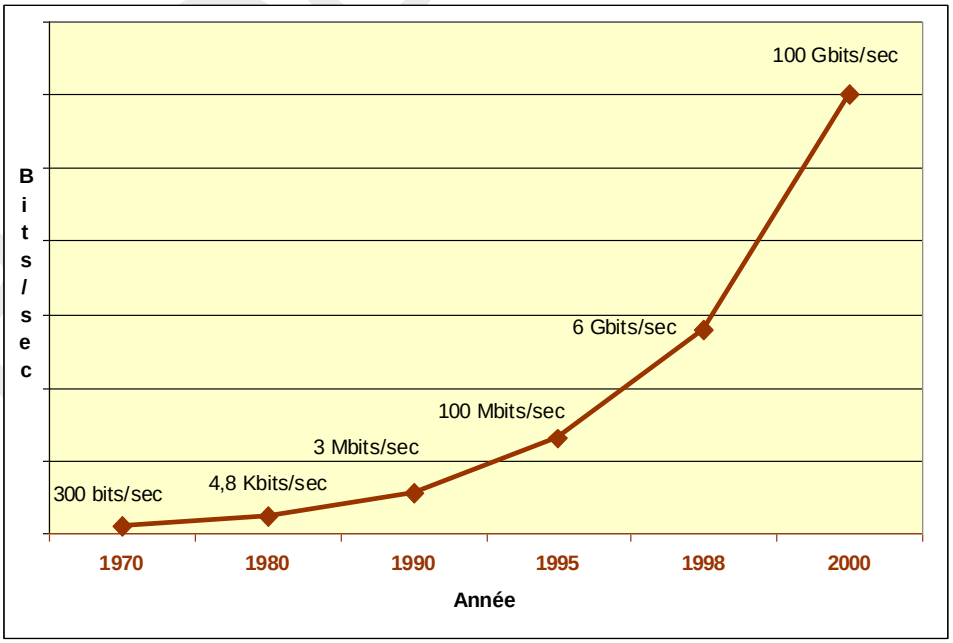
- **Les années 1980-1990**. Le problème de la gestion (collecte, mise à jour, accès, restitution) des données semblait avoir reçu une réponse très satisfaisante. De plus, le développement réussi de la micro-informatique et la baisse des coûts ont poussé les entreprises à s'équiper en ordinateurs et en SGBD pour conquérir une meilleure productivité. Tout le spectre des activités (production, pilotage, décision) a été bouleversé pour s'appuyer sur des systèmes d'information automatisés. De ce fait, on a assisté à une croissance quasi exponentielle du volume des données stockées.

- **Du début des années 1990 à maintenant** : Les entreprises ont amassé des quantités volumineuses de données qui sont maintenant relativement bien gérées, fiabilisées et disponibles à un coût faible. La question qui se pose naturellement à ce stade est de savoir quoi faire de ces données car leur collecte et leur maintenance ont malgré tout un coût, même modeste. Autrement dit, est-il possible de valoriser ces données amassées ? Est-il possible de s'en servir pour prendre des décisions ou éclairer des choix pour l'entreprise ? Est-il possible de dégager un avantage concurrentiel par les connaissances que le l'entreprise peut tirer de ces données ? Ces questions deviennent cruciales quand on estime que le volume des données stockées double tous les ans.



• Figure 1 : évolution de la taille des bases de données

La figure (fig-1) situe cette évolution à la fois en terme de technologies et en terme de volume des données. Aux bases de données relationnelles dont la taille atteignait quelques dizaines de giga-octets, nous sommes passés aux entrepôts de données dont la taille varie de quelques centaines de giga-octets à plusieurs dizaines de téra-octets. Chaque jour de nouveaux records sont annoncés et il ne semble pas se dessiner une limite à la taille des entrepôts en construction. Cette croissance exponentielle va donc se poursuivre sous l'effet particulier d'Internet.



• Figure 2 : Évolution de la vitesse de transmission des réseaux

La vitesse de transmission des réseaux, comme le montre la figure 2, montre une évolution analogue à celle des tailles des entrepôts de données. L'accès distant à

des masses considérables de données pour un coût quasi dérisoire conduit à une duplication des données par copie. En une décennie, nous sommes passés d'une aire où les données étaient rares et où leur regroupement dans une base de données leur donnait une valeur économique intrinsèque vers une aire d'abondance de données où la valeur de celles-ci s'est complètement diluée au profit de l'information ou de la connaissance potentielle qu'elles renferment.

4.2. Le rapport à la clientèle

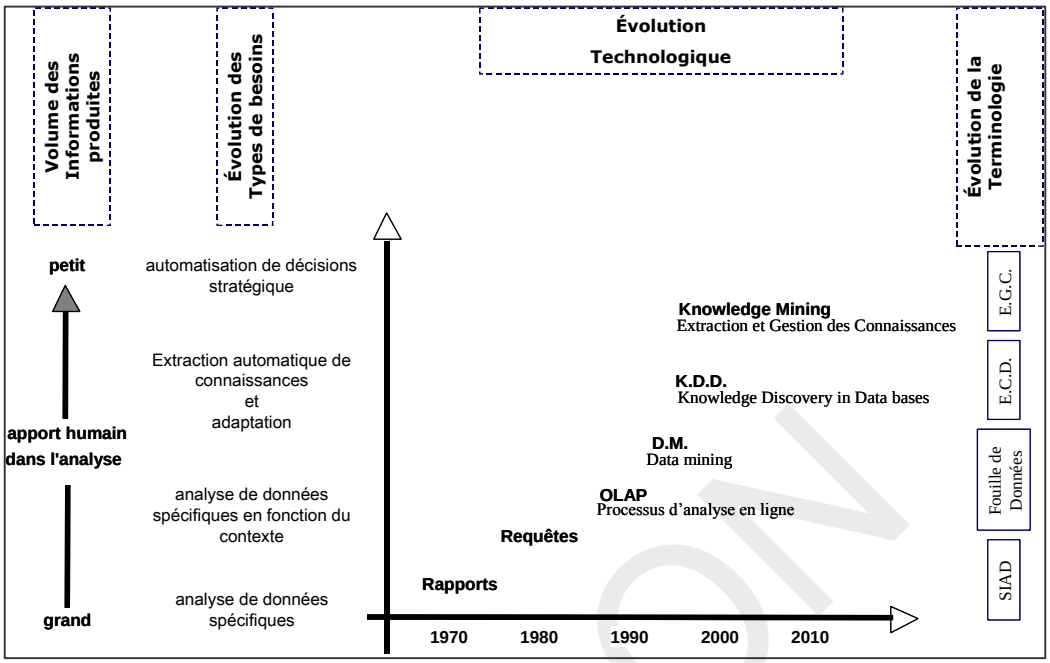
Dans une économie globalisée comme celle qui se développe sous nos yeux, la survie d'une entreprise repose sur l'adaptation de sa réponse à la demande du marché. Nous assistons depuis deux décennies à une transformation du rapport entre l'offre et la demande sur le marché. Par le passé, les entreprises fabriquaient des biens qui étaient ensuite proposés au marché. Nous pouvons qualifier ce modèle d'économie « orientée produit ». De nos jours, la concurrence est plus forte et les clients forts exigeants. Dans un tel environnement, les entreprises sont obligées d'offrir les biens ou les services qui répondent au mieux aux besoins du client, même les plus spécifiques. Nous sommes ainsi passés vers une économie « orientée client ».

Pour mieux répondre à la demande du client, la connaissance de son comportement est décisive, c'est la gestion de la relation client ou *customer relationship management* (CRM).

Les premiers travaux en *data mining* à exploiter les grandes bases de données ont démarré dans le secteur de la gestion de la relation à la clientèle : exploitation des tickets de caisses des supermarchés, exploitation des données de facturation pour les opérateurs de téléphone, etc. Les travaux d'Agrawal sur la découverte de règles d'associations sont parmi les précurseurs : découvrir des associations entre produits achetés pour mieux les disposer sur les rayons, proposer des produits de substitution ou encore mieux cibler une clientèle.

5. Évolution des technologies informatiques de la décision

On aura compris à la lecture des paragraphes qui précèdent que le *data mining* va constituer le moteur essentiel de la décision. Si nous mettons en perspective l'évolution du processus décisionnel et ses technologies, nous obtenons la figure 3. On peut ainsi analyser l'évolution selon 4 canaux.



• Figure 3 : Évolution des technologies informatiques de la décision

- **Le volume d'informations produites** pour la décision s'est considérablement réduit. En effet, on sait mieux analyser les besoins des décideurs pour ne leur fournir que l'information utile à la prise de décision. Dans les années 1970, les décideurs recevaient des piles de listings avec des tableaux chiffrés. De nos jours, ils consultent en temps réel sur Internet les graphiques synthétiques pour la prise de décision.

- **Évolution des types de besoins des utilisateurs**. Le processus d'analyse des données s'est de plus en plus affiné pour prendre en compte le contexte et proposer des modèles pour la décision stratégique.

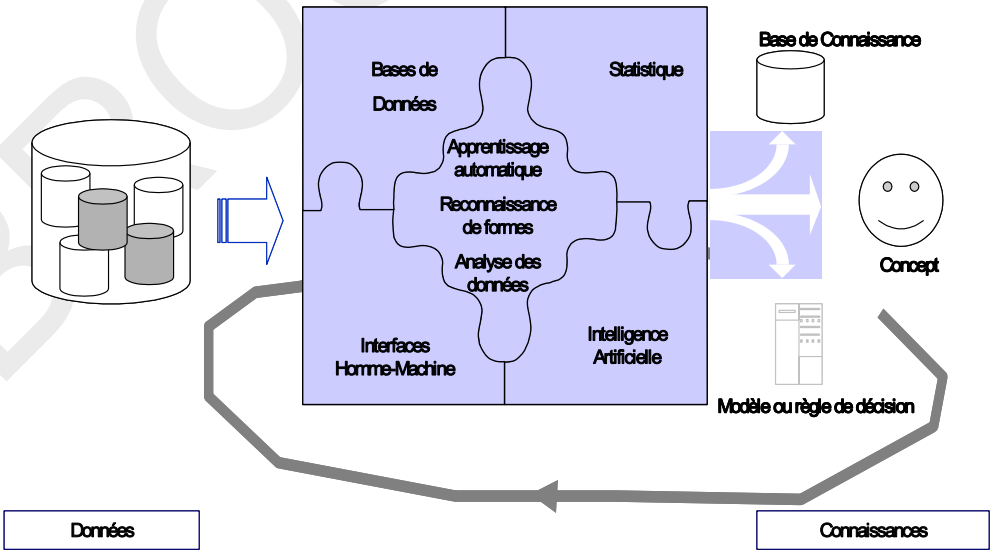
- **Évolution technologique**. Au début des années 1970, on produisait des rapports volumineux sous forme de listings. Vers les années 1980, les moteurs de requêtes comme SQL ont permis de mieux cibler l'information à extraire pour le décideur. L'évolution s'est poursuivie par les systèmes de requêtes et d'analyse en ligne qui opèrent sur des entrepôts de données appelés *on line analytical processing* (OLAP). Cette troisième étape, qui a débuté dans les années 1990 s'est encore dotée d'outils plus sophistiqués comme les méthodes de recherche de règles d'association. Le *data mining* se met en place vers le milieu des années 1990. Des architectures plus complètes et plus complexes ont commencé à apparaître dans les systèmes d'extraction des connaissances à partir des données. Le futur proche semble s'ouvrir à l'extraction et à la gestion des connaissances.

- **Évolution de la terminologie.** Nous sommes passés des systèmes d'information d'aide à la décision (SIAD) vers les systèmes d'extraction et de gestion des connaissances (EGC) travers le *data mining* et l'ECD.

6. L'objet du *data mining*

Les bases de données ou les entrepôts de données, atteignent des volumes de plusieurs tera-octets (1 tera-octet = 10^{12} octets). Les grandes compagnies, comme EDF, France Telecom ou SFR, collectent annuellement plusieurs tera-octets de données relatives aux consommations de leurs clients. Dans un monde de concurrence sévère, chez les grands opérateurs téléphoniques par exemple, la connaissance des clients et de leurs comportements permet de mieux anticiper ceux qui risquent de passer chez un concurrent et de mieux adapter les opérations commerciales pour tenter de les garder. L'une des grandes difficultés est de savoir comment chercher ce profil dans un si grand amas de données. Le *data mining* offre, entre autre, les moyens d'analyse pour chercher s'il existe un profil comportemental typique des clients qui changent de fournisseurs. L'entreprise pourra ainsi repérer plus facilement, parmi ses clients, lesquels ont le profil pour partir vers la concurrence afin de tenter, par des actions commerciales ad hoc, de les garder.

Le *data mining* est un processus qui fait intervenir des méthodes et des outils issus de différents domaines de l'informatique, de la statistique ou de l'intelligence artificielle en vue de découvrir des connaissances utiles.



- Figure 4 : Technologies et modèle général d'ECD

Quand nous parlons d'ECD ou de *data mining*, nous sous-entendons le fait qu'il y ait nécessairement une présence de grandes bases de données. Par ailleurs, les situations où traditionnellement on employait l'expression

« analyse des données », peut-être par un effet de mode, font maintenant plutôt usage de l'expression « *data mining* ».

Comme le montre la figure 4, l'ECD est un processus itératif qui met en œuvre un ensemble de techniques provenant des bases de données, de la statistique, de l'intelligence artificielle, de l'analyse des données, des interfaces de communication homme-machine. L'ECD vise à transformer des données (volumineuses, multiformes, stockées sous différents formats sur des supports pouvant être distribués) en connaissances. Ces connaissances peuvent s'exprimer sous forme d'un concept général qui enrichit le champ sémantique de l'usager par rapport à une question qui le préoccupe. Elles peuvent prendre la forme d'un rapport ou d'un graphique. Elles peuvent s'exprimer comme un modèle mathématique ou logique pour la prise de décision. Les modèles explicites, quelle que soit leur forme, peuvent alimenter un système à base de connaissances ou un système expert.

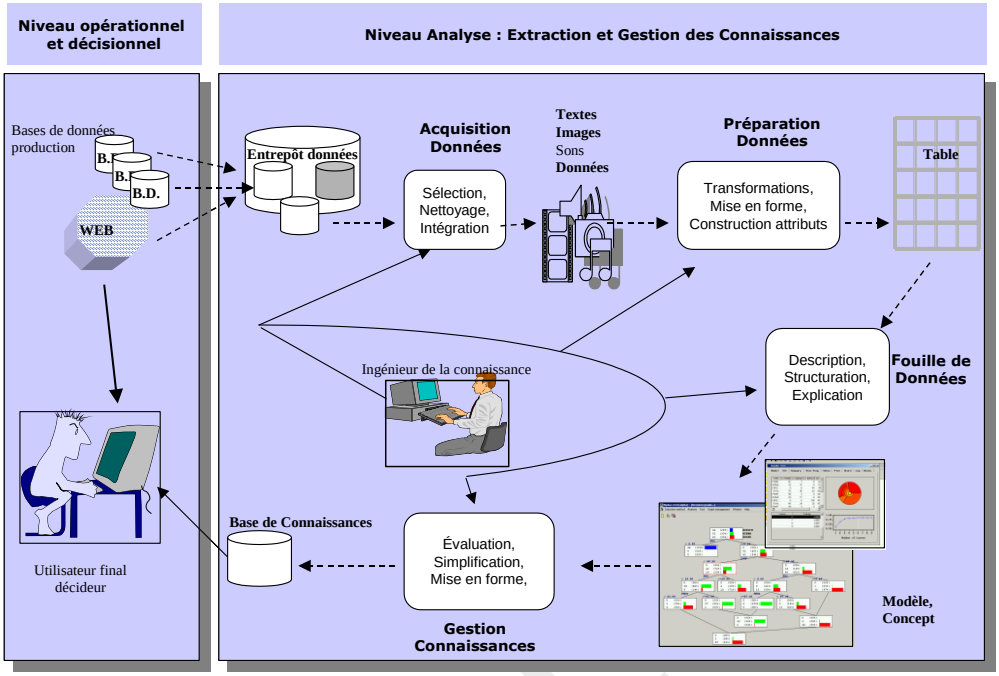
La définition que nous venons de donner nous semble plus générale et mieux adaptée à l'usage du *data mining* moderne que celle proposée initialement par Fayyad en 1996 « *l'extraction de connaissances à partir des données est un processus non trivial d'identification de structures inconnues, valides et potentiellement exploitables dans les bases de données*».

7. Extraction des connaissances à partir des données

L'approche moderne de l'extraction des connaissances à partir des données se veut la plus générale possible. Elle ne privilégie ni une source particulière d'informations (celles-ci peuvent être localement stockées ou distribuées) ni une nature spécifique des données (elles peuvent être structurées en attributs-valeurs, des textes de longueurs variables, des images ou des séquences vidéo). Elle ne se limite pas aux outils d'analyse les plus récents et incorpore explicitement des méthodes pour la préparation des données, pour l'analyse et pour la validation des connaissances produites. Ces méthodes proviennent en majorité de la statistique, de l'analyse des données, de l'apprentissage automatique et de la reconnaissance de formes.

L'ECD est un processus anthropocentré, les connaissances extraites doivent être les plus intelligibles possibles à l'usager. Elles doivent être validées, mises en forme et agencées. Nous allons détailler toutes ces notions et les situer dans le processus général de l'ECD.

L'introduction de l'ECD dans les entreprises est récente. Le rattachement des activités liées à l'ECD n'est pas toujours clair. Selon les cas, elle peut être intégrée au service ou la direction : informatique, organisation, études, statistique, marketing, etc. Comme le montre le schéma de la figure 5, il convient de distinguer le niveau opérationnel et le niveau « analyse » que nous allons décrire.



• Figure 5 : Processus général d'ECD

7.1. Niveau opérationnel et décisionnel

Toutes les actions sont très souvent le résultat d'une décision prise pour répondre à une demande de l'environnement. Ces décisions ou ces actions ne sont bien sûr pas toutes de même importance. Elles peuvent être stratégiques ou de simples actions de routine. Les décisions importantes nécessitent une évaluation qui repose sur des connaissances ou des modèles préétablis. A ce niveau, l'utilisateur cherche à répondre au mieux aux sollicitations de l'environnement. Par exemple, pour un service de vente en ligne, dès qu'un client se connecte sur le site Internet, on cherche à le profiler selon des modèles préétablis pour lui proposer l'offre de service qui est susceptible de l'intéresser. Les modèles utilisés pour profiler les clients sont généralement des programmes informatiques construits en interne ou achetés à des cabinets de conseil spécialisés. Dans le domaine de la gestion de la relation client (CRM), le niveau opérationnel ou décisionnel s'appelle le *front office*. Le *front office* exploite les connaissances qui lui sont fournies par les services études en vue de gérer au mieux la relation avec le client, qu'il soit au guichet, au téléphone ou sur Internet. Les services études sont généralement situés en *back office*. C'est là que s'effectue le processus d'extraction et de gestion des connaissances. Le *front office* est une structure « orientée métier », par exemple un opérateur de téléphone ne met pas les même activités qu'une surface de vente de produits de consommation. En revanche, la structure et les outils qui se trouvent en *back office* peuvent être

rigoureusement les mêmes. C'est le *front office* qui produit et met à jour la base de données de l'entreprise. C'est également lui qui stocke les données historiées dans les entrepôts de données.

7.2. Niveau analyse

C'est le centre des opérations d'extraction des connaissances à partir des données. Les données issues des bases de données de production, en service en *front office*, alimentent les entrepôts de données qui seront utilisées en ECD. Généralement, le processus d'ECD, sous la supervision d'un spécialiste, se déroule en quatre phases : acquisition des données, pré-traitement et mise en forme, fouille de données (*data mining* dans un sens restrictif) et analyse, validation et mise en forme des connaissances.

8. Phase d'acquisition des données

8.1. L'acquisition

Les données peuvent être **localisées sur des sites différents** de celui où s'effectue l'ECD. Elles peuvent être **stockées selon des architectures variées** : dans des bases de données relationnelles, dans des entrepôts de données, sur le *web* ou dans des banques de données spécialisées (images, bibliothèques ou librairies numériques, base de données génomiques). Elles peuvent être **structurées ou non** selon différents types : données tabulaires ou textuelles, images, sons ou séquences vidéo. En ECD, l'analyste, qu'il soit ingénieur de la connaissance ou statisticien, doit avoir un problème relativement bien délimité. Il ne se lance pas dans l'ECD sans avoir une certaine idée des objectifs de son opération et des moyens informationnels et technologiques dont il dispose. Par exemple, il souhaite comprendre pourquoi certains de ses clients se sont tournés vers une entreprise concurrente ou il cherche à mieux connaître son activité selon différents critères. Toutes les données disponibles et accessibles au niveau de l'entrepôt ne sont certainement pas utiles dans leur intégralité pour traiter son problème particulier. Il ne viendrait à l'esprit d'aucun spécialiste en *data mining* de télécharger tout le contenu du *web* (évalué à plusieurs milliards de pages) pour en extraire des connaissances, d'autant plus que le contenu du *web* change quasiment à tout instant.

La phase d'acquisition vise ainsi à cibler, même de façon grossière, l'espace des données qui va être exploré, le spécialiste du *data mining* agit ainsi un peu à l'image du géologue qui définit des zones de prospection, étant persuadé que certaines régions seront probablement vite abandonnées car elles ne recèlent aucun ou peu de minerais. L'acquisition met en œuvre des requêtes ad hoc pour rapatrier les données potentiellement utiles selon le point de vue de l'expert. Le processus d'ECD n'est pas linéaire car il arrive aussi que l'on revienne, après analyse, rechercher de nouvelles données. La

phase d'acquisition nécessite le recours à des moteurs de recherche de données. Cette phase peut passer par les moteurs de requêtes des bases de données comme le langage SQL. L'acquisition peut aussi se faire à travers des outils de requêtes plus spécifiques aux données non structurées comme les données textuelles, les images ou le *web*, faisant pour cela appel à des moteurs de recherche d'informations et d'images auxquelles ils accèdent par le contenu.

Cette phase d'acquisition sert généralement à nettoyer les données qui sont rapatriées. Par exemple, si l'un des attributs retenus s'avère au moment du rapatriement peu ou mal renseigné, on peut le laisser tomber tout de suite. On peut également explicitement chercher à limiter le nombre d'enregistrements que l'on souhaite traiter. On construit alors un filtre idoine. Il peut être de nature statistique, par exemple, un échantillonnage au 1/1000 selon une procédure de tirage aléatoire simple.

A l'issue de la phase d'acquisition, l'analyste est, *a priori*, en possession d'un stock de données contenant potentiellement l'information ou la connaissance recherchée. Il convient de préciser que l'ECD s'effectue toujours sur un échantillon du monde réel même s'il atteint plusieurs téra-octets. N'oublions pas que l'ECD ne travaille que sur des données relatives à des événements passés, très souvent, nous n'avons jamais accès à la totalité des données liés à un phénomène réel.

8.2. Pré-traitement des données

Les données issues des entrepôts ne sont pas nécessairement toutes exploitables par des techniques de fouille de données. En effet, la plus part des techniques que nous utilisons ne traitent que des tableaux de données numériques rangées sous forme lignes/colonnes. Certaines méthodes sont plus contraignantes que d'autres. Elles peuvent par exemple exiger des données binaires, comme c'est le cas des premières techniques de recherche de règles d'association.

Les données acquises depuis l'entrepôt peuvent être de types différents. On peut y trouver des textes de longueur variables, des images, des enregistrements quantitatifs ou des séquences vidéo.

La préparation consiste à homogénéiser les données et à les disposer en tableau lignes/colonne. Formellement, chaque ligne/colonne peut être considérée comme un objet vecteur ayant un nombre fixe de composantes. Ce vecteur ligne/colonne sera vu comme un objet mathématique que l'on pourra manipuler selon qu'il possède ou non certaines propriétés. Par exemple, si tous les vecteurs lignes sont des points de l'espace euclidien à p dimensions, on pourra faire appel aux techniques de *data mining* basées sur l'algèbre linéaire.

Gardons à l'esprit qu'*in fine* cette transformation doit déboucher sur un tableau ligne/colonne car il s'agit presque toujours de la structure la mieux adaptée à l'exploitation des données. Précisons que dans certaines situations, les données arrivent déjà sous une forme appropriée et qu'il n'est alors plus nécessaire de les modifier. Dans d'autres cas, elles sont dans une structure tabulaire mais exigent une transformation telle qu'un centrage par rapport à la moyenne ou une normalisation. En fait, le pré-traitement est un acte de modélisation d'expert. Si l'expert ne définit pas les bonnes transformations ou les bons attributs, il ne verra alors rien dans ses données. L'expert devra par conséquent choisir un canevas pour représenter ses données et éventuellement effectuer une série de transformations pour obtenir des données adaptées aux méthodes d'exploitation.

Les principales opérations de préparation peuvent être listées comme suit :

- **Sélection de ligne/colonne**. Elle s'effectue sur des données qui sont déjà sous forme tabulaire. Il s'agit ensuite de définir un filtre qui permet de sélectionner un sous-ensemble de lignes ou de colonnes. L'objectif est soit de réduire le nombre de données soit de sélectionner les lignes ou colonnes les plus pertinentes par rapport aux préoccupations de l'utilisateur. Les techniques mises en œuvre dans ce but relèvent des méthodes statistiques d'échantillonnage, de sélection d'instances ou de sélection d'attributs. Cette sélection peut également s'effectuer selon des conditions exprimées par l'utilisateur. Par exemple, il peut ne garder que les attributs dont la moyenne est supérieure à un seuil donné ou ne conserver que les attributs qui ont un lien statistique significatif avec un attribut particulier. Ce lien sera évalué à l'aide d'une mesure d'association comme le khi-2 de Pearson ou le gain informationnel. La sélection d'attributs est en train de devenir l'un des sujets majeurs de la recherche en *data mining*.

- **Le traitement des données manquantes ou aberrantes**. Certaines données peuvent être absentes et gêner ainsi l'analyse. Il faut donc définir des règles pour gérer ou pour remplacer ces données manquantes. De nombreuses solutions sont proposées, comme le remplacement, dans le cas des données numériques continues, de toute donnée manquante par le mode de la distribution statistique (la valeur la plus fréquente) de l'attribut concerné, si ce mode existe. On peut également chercher à estimer ces valeurs manquantes par des méthodes d'induction comme la régression, les réseaux de neurones simples ou multicouches, ou les graphes d'induction. Pour le traitement des données aberrantes, il faut d'abord repérer ces dernières au moyen d'une règle préétablie. Par exemple, toutes les données numériques dont la valeur sur un attribut donné s'écarte de la valeur moyenne plus deux fois l'écart-type, pourraient être considérées comme des données possiblement aberrantes et qu'il conviendrait de traiter.

- **Les transformations d'attributs.** Il s'agit de transformer un attribut A en une autre variable A' qui serait, selon les objectifs de l'étude, plus appropriée. Différentes méthodes sont pratiquées comme la discrétisation qui consiste à transformer des attributs continus en découpant le domaine de valeurs de ces attributs en intervalles afin d'obtenir des attributs qualitatifs. Il existe à cet effet pléthore de méthodes de discrétisation : supervisées ou non, à intervalles de tailles identiques, ou à intervalles à effectifs constants. On peut également centrer par rapport à la moyenne et réduire par l'écart type les valeurs des variables continues. Ce traitement leur confère certaines propriétés mathématiques intéressantes lors de la mise en œuvre de méthodes d'analyse des données multidimensionnelles.

- **La construction d'agrégats.** Dans certaines situations particulières, il peut s'avérer que des agrégats d'attributs soient très importants pour la tâche d'analyse. Un agrégat d'attribut est un nouvel attribut obtenu selon une transformation précise. Par exemple, le prix au mètre-carré d'un appartement, défini par le rapport entre le prix de l'appartement et la surface totale de l'appartement, fournit une indication assez pertinente pour comparer les appartements ou les quartiers dans les bases de données spatiales. On peut imaginer une multitude de façons d'obtenir des agrégats. Parmi les méthodes de construction d'agrégats les plus utilisées, les méthodes factorielles telles que l'analyse en composantes principales (ACP) ou l'analyse des correspondances multiples (ACM) sont largement utilisées. Il n'y a pas de règles précises pour dire que tel agrégat est meilleur qu'un autre. C'est la connaissance du domaine par l'expert qui peut guider dans la définition des bons agrégats. Un agrégat peut être évalué *a posteriori* selon des règles analogues à celles utilisées en sélection d'attribut décrites plus haut.

- **Le traitement des données complexes.** Toutes les méthodes de pré-traitement citées précédemment opèrent sur des tableaux de données lignes/colonnes. Or il arrive que nous travaillions sur des données non structurées sous forme de tableaux. Par exemple, en fouille de texte, nous disposons d'un ensemble de textes de longueurs variées qu'il convient de ramener à une forme tabulaire. L'une des techniques les plus simples consiste à recenser l'ensemble des mots de tout le corpus et ensuite de calculer, pour chaque texte représenté par une ligne ou une colonne, la fréquence de chacun de ces mots. On obtient ainsi un tableau de comptage. Mais le codage des textes fait généralement appel à des procédures plus élaborées qui s'appuient sur la linguistique : lemmatisation, suppression des mots vides, thesaurus ou ontologies du domaine. La préparation des données images, et *a fortiori* vidéo, est encore plus ardue pour le non-spécialiste car il faut définir la liste des attributs qui décrivent l'image. Par exemple, on utilise les caractéristiques de l'histogramme des niveaux de gris, les attributs de texture, etc. La fouille sur des images peut nécessiter d'autres transformations en amont qui relèvent plus du traitement de

l'image que de l'ECD. L'extraction de connaissances à partir de données complexes est d'ailleurs un domaine en pleine croissance.

9. Phase de fouille de données

La fouille de données concerne le *data mining* dans son sens restreint et est au cœur du processus d'ECD. Cette phase fait appel à de multiples méthodes issues de la statistique, de l'apprentissage automatique, de la reconnaissance de formes ou de la visualisation. Les méthodes de *data mining* permettent de découvrir ce que contiennent les données comme informations ou modèles utiles. Si nous essayons de classifier les méthodes de fouille de données utilisées, trois catégories se distinguent :

- Les méthodes de visualisation et de description.
- Les méthodes de classification et de structuration.
- Les méthodes d'explication et de prédiction.

Chacune de ces familles de méthodes comporte plusieurs techniques appropriées aux différents types de tableaux de données. Certaines sont mieux adaptées à des données numériques continues alors que d'autres sont plus généralement dédiées aux traitements de tableaux de données qualitatives. Nous allons donner à présent un aperçu général sur les principales méthodes.

9.1. Les méthodes de visualisation et de description

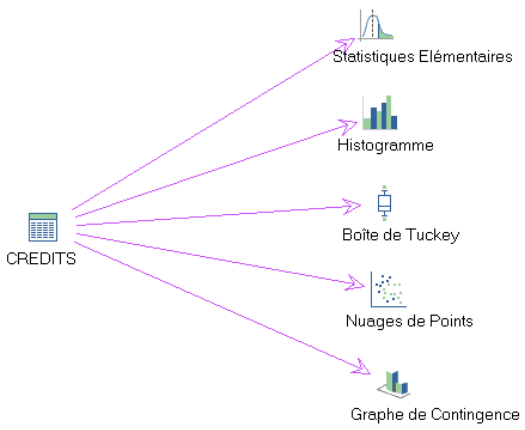
L'objectif de ces méthodes est de permettre à l'analyste d'avoir une compréhension synthétique de l'ensemble de ses données. Il s'agit donc principalement d'outils de synthèse d'information. Cette synthèse peut s'exprimer par des indicateurs statistiques. Par exemple, pour des attributs quantitatifs, les indicateurs les plus utilisés sont la moyenne, l'écart-type, le mode et la médiane. Pour des attributs qualitatifs, on associe généralement la distribution selon les modalités de l'attribut. Ces indicateurs statistiques, qu'ils soient descriptifs de la tendance centrale, des positions ou de la dispersion nous renseignent pleinement sur une caractéristique particulière de la population. Ils sont généralement représentés par des graphiques, car plus faciles à interpréter, comme les boites de Tuckey, les distributions (densités ou fonctions de répartition), les nuages de points. On trouve dans les logiciels de *data mining* une kyrielle de formes géométriques et de styles de présentation de ces concepts.

La description et la visualisation peuvent être mono ou multidimensionnelles. Pour l'essentiel, il s'agit de rendre visible des objets ou des concepts qui se trouvent dans des espaces de description trop riches.

Considérons le cas d'une banque qui enregistre l'ensemble des transactions commerciales, soit plusieurs millions par an, qu'elle réalise avec ses clients : retraits, prêts, dépôts, etc. Cette banque souhaite développer un plan d'étude visant à mieux connaître sa clientèle.

Imaginons que parmi les facteurs qui intéressent la banque dans le cadre d'une première approche du problème figurent l'âge des clients, les montants des crédits qui leur sont alloués, les villes de résidences des clients, les destinations des prêts (acquisition d'une maison, d'une voiture ou d'un équipement domestique) et la période de l'année (par exemple le numéro du trimestre).

Le schéma de la figure 6 représente un processus de *data mining* orienté vers la visualisation et la description. Pour simplifier la présentation, le tableau des données (CREDITS) ne contient que 1000 clients. Les traitements qui figurent aux extrémités des flèches synthétiseront les données des clients de cette banque selon différentes caractéristiques, qu'elles soient numériques ou graphiques.



• Figure 6 : traitement de description et de visualisation

Nous reproduisons ci-après, à titre illustratif, quelques résultats.

Les représentations graphiques

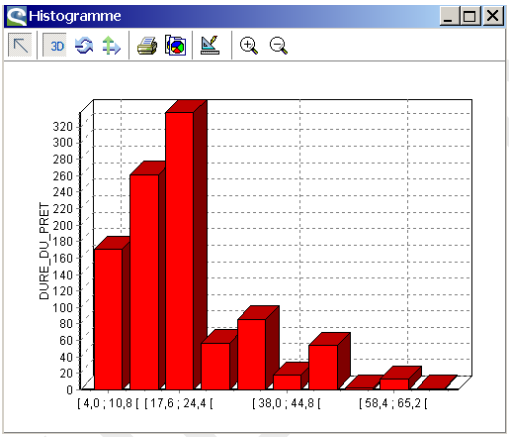
- **Statistiques élémentaires**. Elles sont calculées sur deux attributs quantitatifs : *l'âge du client* et *le montant du crédit* qui lui a été alloué.

Visualisation Statistiques Elémentaires

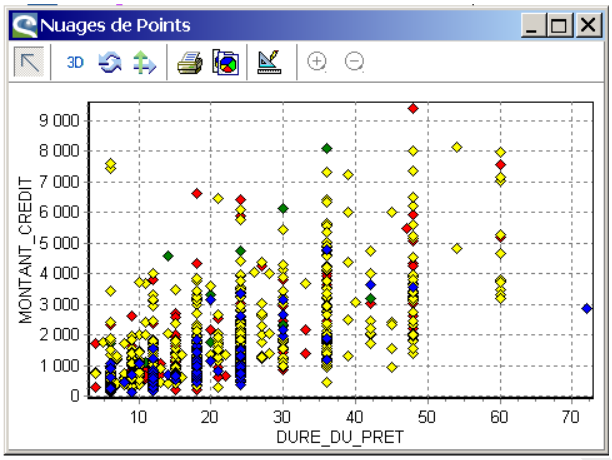
Individus sélectionnés : 1000

	MONTANT_CREDIT	AGE
Moyenne	1668,342	35,546
Ecart Type	1439,596	11,375
Skewness	1,944	1,018
Sign. Skw	0,000	0,000
Kurtosis	4,293	0,596
Sign. Krt	0,000	0,000
Médiane	1183,200	33,000

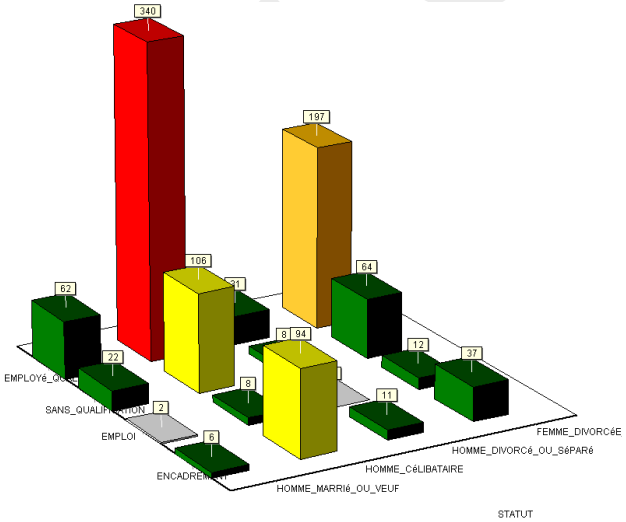
- **Histogramme**. Il est construit sur l'attribut « *durée du prêt* ».



- **Nuage de points**. Il croise les attributs *durée* et *montant du crédit*. Il s'agit là d'une description bidimensionnelle. Elle permet de visualiser, entre autre, si un lien existe ou pas entre ces deux facteurs. Sur ce graphique on peut déjà noter l'existence de points marginaux (des points isolés) sur lesquels on peut s'interroger : est-ce qu'il s'agit de points aberrants ou de points atypiques ? Une liaison faible existe entre *le montant* et *la durée*. En bref, ce type de graphique peut être enrichi en faisant figurer par exemple, pour chaque point, une couleur différente selon que le client est un homme ou une femme. On peut complexifier encore davantage le graphique mais il ne faut pas perdre de vue l'objectif d'une telle représentation : comprendre en un coup d'œil ce qu'il y a d'informatif dans les données.



- **Graphe de contingence**. Il donne les effectifs croisés entre le type d'activité (employé, sans qualification, avec qualification, encadrement) et le statut familial (homme ou femme, célibataire, marié(e), divorcé(e) ou veuf(ve)).

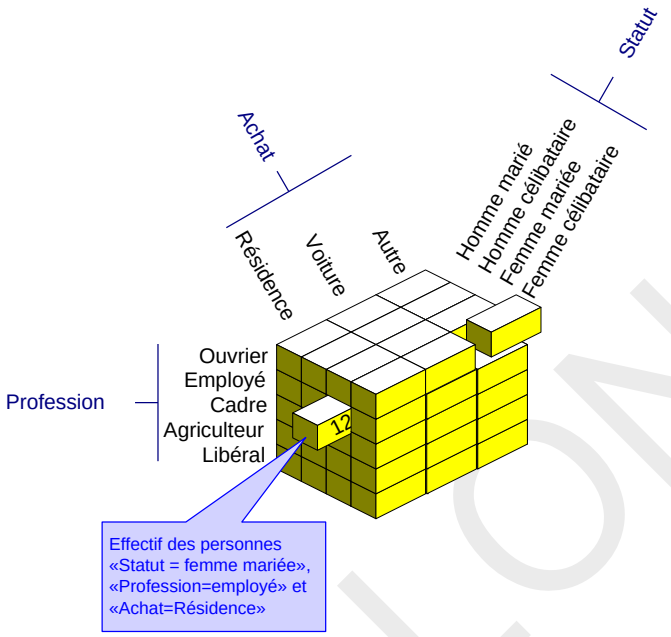


La hauteur de la barre est proportionnelle à l'effectif qui se trouve au croisement de deux modalités. Le concept des tableaux de contingence a été relativement bien exploité dans le domaine de la visualisation des données multidimensionnelles. En effet, les systèmes OLAP exploitent essentiellement des tableaux de contingence tri-dimensionnels appelés des cubes. A l'intersection d'un ensemble, ou *tuple*, de trois modalités, appelées dans la terminologie OLAP « dimensions », se trouve un indicateur comme l'effectif ou la moyenne d'un quatrième attribut.

Les cubes de données

Les cubes de données introduits avec les systèmes OLAP fournissent des tableaux de contingence multidimensionnels, généralement tri-

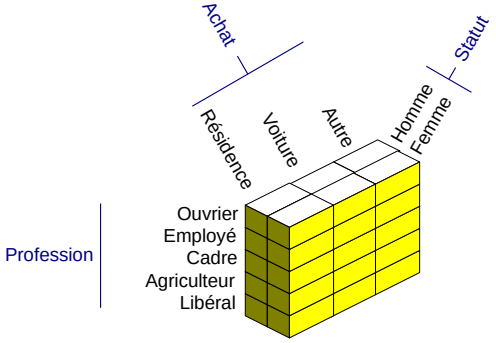
dimensionnels, sur lesquels ont été définies des opérations facilitant l'exploration des données.



• Figure 7 : cube de données

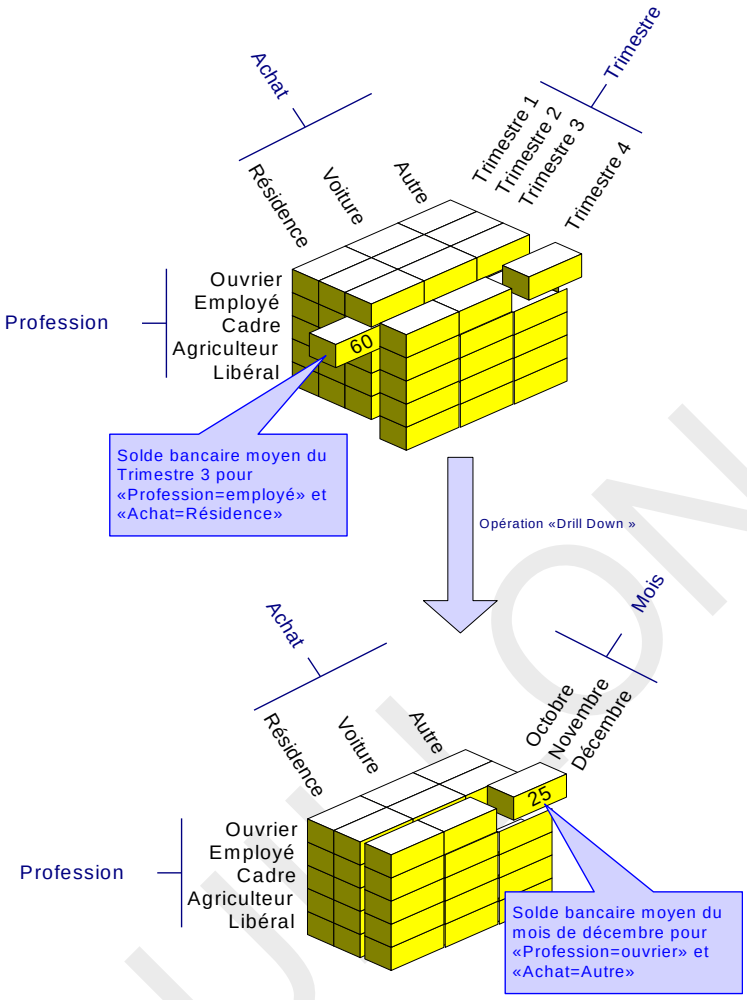
La figure 7 montre un cube de données. Les attributs *statut*, *achat* et *profession* en sont les dimensions. Chaque case, croisement de 3 modalités, figure l'agrégat. Dans ce cas, il s'agit de l'effectif mais on aurait pu imaginer afficher le montant moyen des crédits demandés par les individus de cette case ou la distribution des montants demandés.

On peut décider d'explorer une sous population particulière. Par exemple, par une opération appelée *roll up* dans la terminologie OLAP, on regroupe les modalités *mariés* et *célibataires* afin d'avoir hommes d'un côté et les femmes de l'autre.



• Figure 8 : Opération *roll up* sur un cube de données

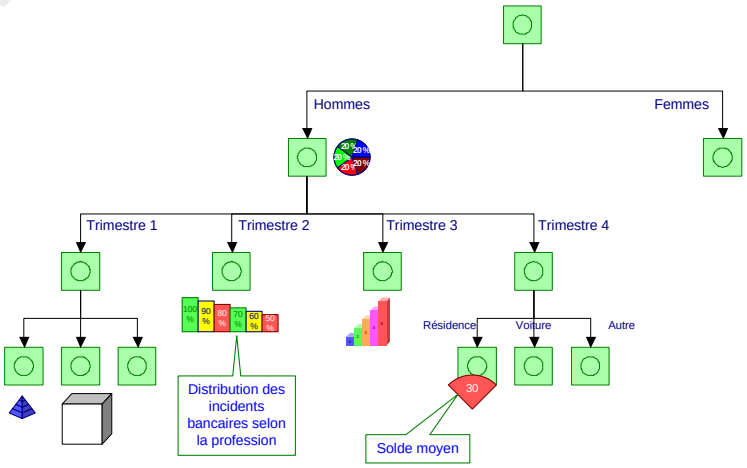
On peut également, par une opération de *drill down* approfondir l'analyse sur un niveau de détail plus fin. Par exemple, sur la figure 9, on a détaillé le solde moyen mensuel au lieu du trimestriel.



• Figure 9 : Opération *drill down* sur un cube de données

Les arbres

On peut imaginer d'autres opérations sur des tableaux à p dimensions. Dans ce cas, il faudra utiliser une représentation par un arbre n-aire.



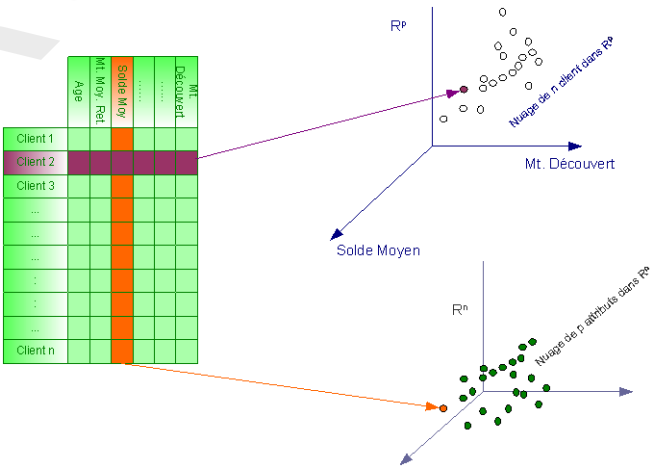
• Figure 10 : arbre de description

Sur un arbre, chaque nœud correspond à une cellule de l'hyper-cube. Nous pouvons représenter qu'une partie de l'arbre qui nous intéresse. A l'intérieur de chaque feuille ou sommet de l'arbre il est possible de représenter des informations statistiques numériques ou graphiques sur des agrégats.

L'analyse des données multidimensionnelles

Il arrive très souvent qu'une description selon un ou deux attributs ne soit pas satisfaisante et que l'utilisateur souhaite prendre en compte simultanément la totalité des attributs. Nos possibilités visuelles ne nous permettent malheureusement pas de voir des objets qui sont dans des espaces à plus de 3 dimensions. L'analyse des données multidimensionnelles nous fournit le moyen d'accéder à cette description et de visualiser au mieux les données sous leur forme résumée.

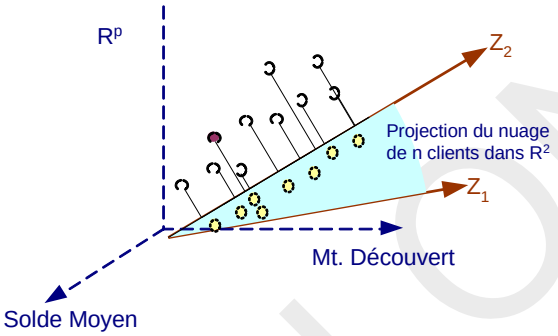
Les méthodes d'analyse des données multidimensionnelles opèrent sur des tableaux numériques. Il peut s'agir d'un tableau de p mesures prises sur un ensemble de n individus. Par exemple, sur l'ensemble des clients d'une banque, nous disposons de l'âge, du montant moyen des retraits par mois, du montant du découvert maximum constaté, du solde moyen, etc. Ainsi, un client est alors caractérisé par un vecteur à p composantes. Moyennant de petites précautions pour rendre homogènes ces données, on peut considérer chaque client comme un point de l'espace euclidien $\mathbf{R}^p$. L'ensemble des n clients forme alors un nuage de points plongé dans $\mathbf{R}^p$. On peut tout aussi bien imaginer la situation duale où nous avons un nuage de p attributs plongés dans l'espace euclidien $\mathbf{R}^n$ des individus.



• Figure 11 : Nuage de points dans un espace multidimensionnel

Les deux nuages de points ne sont pas visibles à l'œil nu dès lors que p et n dépassent 3. Pour contourner ce handicap, nous devons les projeter dans des espaces de faibles dimensions : une droite, un plan ou un espace à trois

dimensions. Cette opération de projection est analogue à la prise d'une photo par une caméra : des points de l'espace à trois dimensions sont projetés sur la plaque photographique qui est à deux dimensions. Indépendamment des considérations artistiques, les différentes prises de vues d'une scène ne révèlent pas la même information. Si nous cherchons un résumé qui se veut être le plus fidèle possible de la réalité, nous devons rechercher le sous-espace de dimension 1, 2 ou 3 qui conserve au mieux les proximités originales entre les points. Sur la figure suivante, nous avons ainsi cherché le « meilleur » plan de projection des points clients.



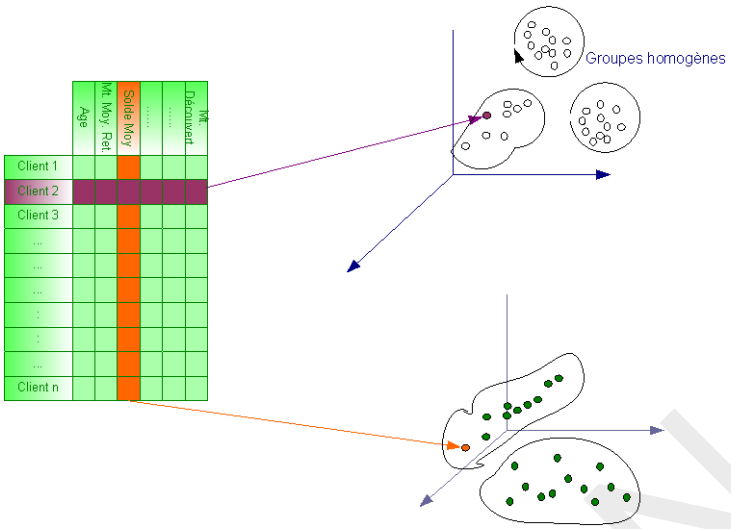
• Figure 12 : Projection des points sur le meilleur plan factoriel

Les axes Z_1 et Z_2 sont appelés des facteurs. Ils fournissent une représentation qui déforme le moins possible les distances originales entre points. Les principales méthodes utilisées pour extraire ces facteurs sont l'analyse en composantes principales et l'analyse des correspondances abondamment décrite dans les manuelles d'analyse des données dont plusieurs sont fournis en bibliographie. Elles s'appuient essentiellement sur les résultats et les propriétés de l'algèbre linéaire.

Dans le domaine de la visualisation des données, les recherches se sont grandement intensifiées. De nombreux ouvrages cités en bibliographie sont également disponibles.

9.2. Les méthodes de structuration et de classification

En ECD, sans doute plus qu'ailleurs, nous avons affaire à une profusion de données. Décrire ces données s'avère parfois difficile à cause de cette volumétrie. L'utilisateur cherche souvent à identifier des groupes d'objets semblables au sens d'une métrique donnée. Ces groupes peuvent par exemple correspondre à une réalité ou à des concepts particuliers.



• Figure 13 : Méthodes de structuration

Les lignes ou les colonnes du tableau sont vues comme des points d'un espace multidimensionnel qui n'a pas obligatoirement une structure d'espace vectoriel. Les méthodes de structuration ont pour objet de repérer ces structures de groupe invisibles à l'œil nu.

Par exemple, dans le domaine du marketing, il est impensable de construire un message spécifique pour chaque client potentiel. Un service de marketing va chercher à identifier des groupes d'individus semblables selon différents critères de telle sorte que la campagne soit ciblée sur quelques groupes.

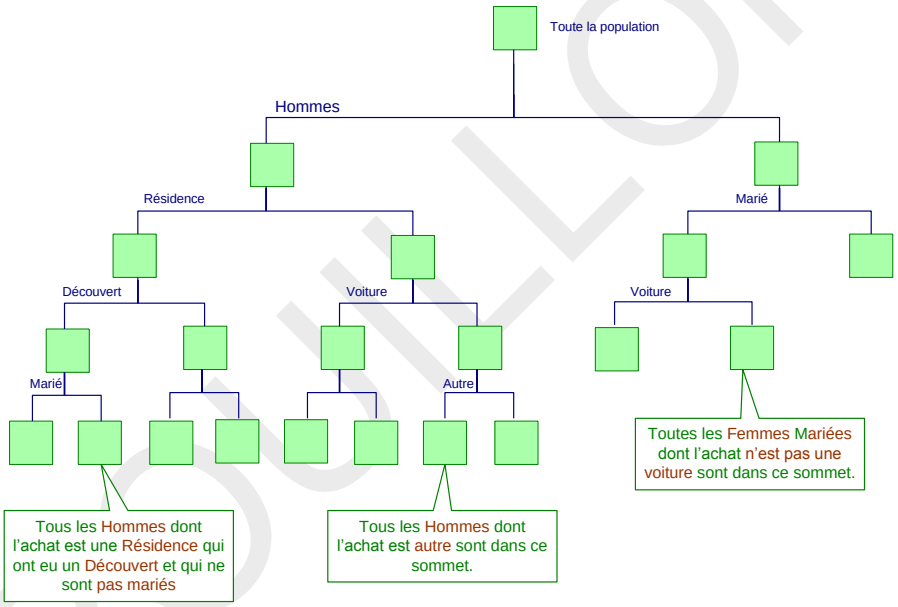
Dans la réalité, l'homme a souvent beaucoup de mal à mémoriser de façon individualisée un ensemble d'objets, surtout quand ils sont en très grand nombre. Par exemple, mémoriser toutes les espèces végétales ou animales est une tâche extrêmement laborieuse, voire impossible, pour un humain. L'homme préfère généralement catégoriser ces objets en classes en fonction de certaines propriétés communes ou en fonction d'un critère donné. Ces classes ou ces catégories d'objets sont ensuite nommées. Par exemple, le monde animal est structuré en groupes : vertébrés ou invertébrés, mammifères ou non, etc. Ainsi, toutes les espèces sont ventilées en fonction de la présence ou non de certains attributs communs. Il est évident que cette classification est un pur produit intellectuel et qu'elle n'a pas d'existence propre. Ces classes sont en fait des objets appelés parfois « concepts » car ils n'existent pas en tant que tels dans la nature.

Les techniques employées pour des opérations de classification relèvent de ce que nous appelons *l'apprentissage non supervisé*. Nous parlons d'apprentissage non supervisé car l'utilisateur ne sait pas *a priori* quelles classes, groupes ou catégories il va obtenir. Ce mode d'apprentissage est également appelé « apprentissage sans professeur ».

Les techniques employées sont des méthodes de classification automatique (*cluster analysis*), appelées aussi « classifications conceptuelles » ou « méthodes de taxinomie ». Elles sont exposées dans de nombreux ouvrages de reconnaissances de formes, d'analyse des données et d'intelligence artificielle.

Les principales techniques se répartissent en trois groupes :

- Les **méthodes monothétiques** dont l'objet est la recherche de partitions sur l'ensemble des objets à classer, telles que sur chaque classe, l'un des attributs X_i soit constant ou de très faible variance. Par exemple, dans la classe des vertébrés, toutes les espèces ont en commun la présence de vertèbres. Dans cette catégorie de méthodes, on peut citer la *segmentation* de Williams et Lambert.

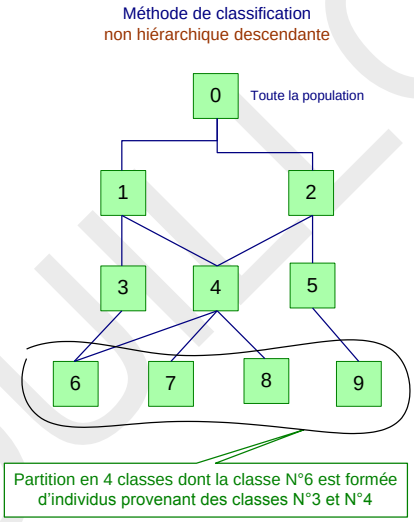


• Figure 14 : méthode de segmentation pour la classification monothétique

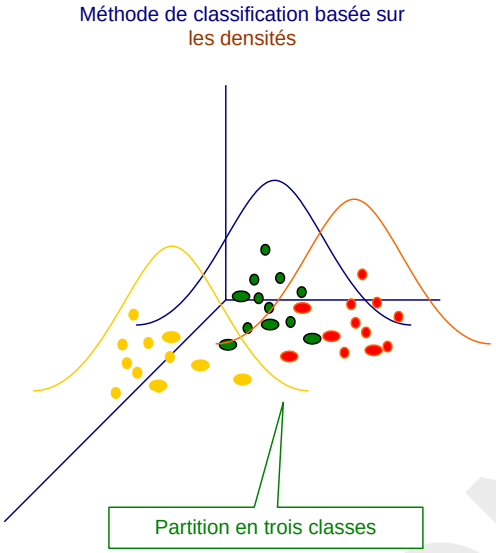
Cette méthode de classification suppose que le tableau des données soit binaire (absence/présence d'un attribut). Le critère d'homogénéité pour caractériser les groupes est basé sur le lien du khi-2. L'algorithme segmente selon un attribut si les deux sous-groupes générés à partir d'un attribut binaire sont les plus homogènes au sens de ce critère. Le processus est réitéré sur chaque nœud de manière indépendante jusqu'à épuisement des attributs ou jusqu'à la satisfaction d'un critère d'arrêt généralement fixé par l'utilisateur. Le résultat est une hiérarchie de partitions où la racine de l'arbre contient la partition grossière.

- Les **méthodes polythétiques**, de loin les plus nombreuses. Ces méthodes recherchent des partitions dans lesquelles les éléments d'une même classe ont, entre eux, une certaine ressemblance, et des

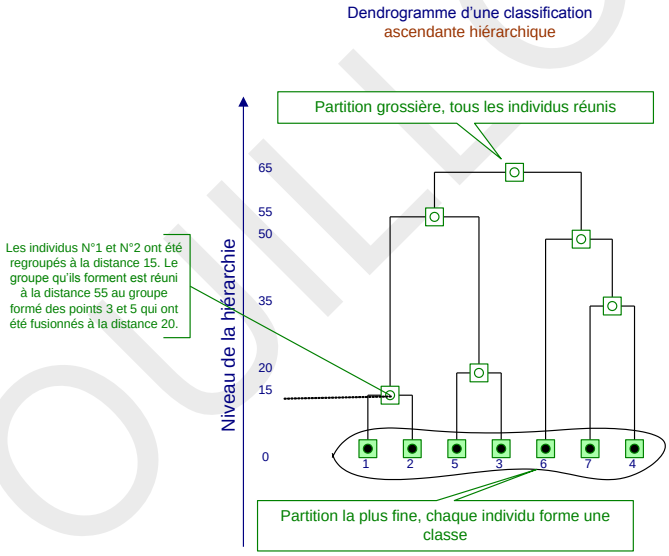
éléments appartenant à des classes différentes d'une même partition qui doivent être les plus dissemblables possibles au sens d'un certain critère préétabli. La ressemblance doit prendre en compte la totalité des attributs descriptifs. On distingue les méthodes selon qu'elles conduisent à une hiérarchie de partitions emboîtées ou non, ou une partition à nombre de groupes prédéterminé ou non. Parmi les techniques fréquemment employées, on trouve les méthodes de classification hiérarchique, les nuées dynamiques proposée par Diday, la classification non hiérarchique descendante proposée par Fages,... On peut également incorporer les algorithmes développés dans les domaines de l'intelligence artificielle comme l'algorithme Etoile proposé par Michalski qui s'inspire de l'algorithme des nuées dynamiques ou les algorithmes COBWEB, et AUTOCLASS. Les figures suivantes donnent les représentations graphiques que l'on rencontre le plus souvent.



• Figure 15 : Classification non hiérarchique descendante

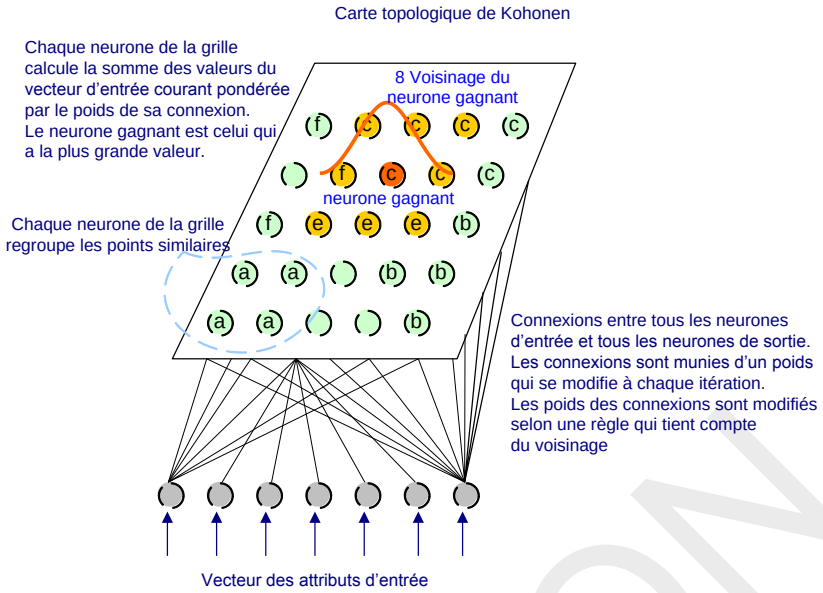


• Figure 16 : Classification basée sur les centroïdes



• Figure 17 : Classification ascendante hiérarchique.

- Les **méthodes basées sur les réseaux de neurones**. Le processus d'apprentissage est incrémental, c'est-à-dire que les objets sont affectés séquentiellement à des groupes en fonction de leur proximité. Nous retrouvons, dans cette catégorie de méthodes, les techniques dites des cartes topologiques de Kohonen et les techniques basées sur la résonance adaptative de Grossberg et Carpenter.



• Figure 18 : Carte topologique de Kohonen

Le choix des méthodes dépend essentiellement des objectifs du chercheur. Il n'existe pas de principes universels conduisant au meilleur choix possible. Généralement, on utilise plusieurs techniques et on compare leurs résultats. Si l'on constate une certaine stabilité, cela signifie que nous sommes en présence de structures réelles. Bien sûr, si les résultats présentent un sens concret du point de vue de l'utilisateur, c'est en soit une condition qui peut être suffisante pour retenir la catégorisation proposée.

9.3. Les méthodes d'explication et de prédiction

Ces méthodes ont pour objectif de rechercher à partir des données disponibles un modèle explicatif ou prédictif entre, d'une part, un attribut particulier à prédire et, d'autre part, des attributs prédictifs. Dans le cas où un tel modèle serait produit et qu'il s'avérait valide, il pourrait alors être utilisé à des fins de prédiction.

Considérons le cas d'un médecin qui s'intéresse à la nature d'une affection dont il veut connaître la nature cancéreuse ou non. Imaginons qu'il souhaite construire une règle lui permettant de prévoir, à l'avance sur la base d'examens cliniques simples, la nature cancéreuse ou bénigne de l'affection. Pour cela, il peut procéder par apprentissage à partir de données. Cela consiste, pour lui, à recueillir des informations sur des patients déjà traités pour cette affection et dont il sait si elle a été cancéreuse ou bénigne. Sur la base de ce corpus qu'on appellera « échantillon d'apprentissage », il mettra en œuvre une méthode d'apprentissage qui l'aiderait à bâtir son modèle d'identification. Dans ce contexte on parle d'apprentissage supervisé car l'attribut à prédire est déjà préétabli. Il s'agit alors de mettre au point un processus permettant de le reconstituer de façon automatique à partir des autres attributs.

En apprentissage supervisé, il y a, d'une part, une phase « inductive » consistant à développer les règles d'identification à partir d'exemples particuliers et, d'autre part, une phase « prédictive » visant à utiliser ces règles pour identifier de nouvelles instances. Cependant les méthodes de prédiction ne procèdent pas toutes ainsi. Les méthodes de prédiction à base d'instance comme les *k plus proches voisins* n'établissent pas une liaison fonctionnelle entre l'attribut à prédire et les valeurs des attributs prédictifs.

Il existe une multitude de méthodes d'explication et ou de prédiction développées dans différents contextes. En dehors des méthodes à base d'instance, nous allons présenter synthétiquement les principales familles de méthodes d'explication et de prédiction, le lecteur intéressé par les méthodes à base d'instance pourra se référer aux ouvrages de reconnaissance de formes comme celui de Duda et Hart donné en bibliographie.

On considère une population d'individus ou d'objets notée Ω concernés par le problème d'apprentissage. A cette population est associé un attribut particulier appelé « attribut endogène » noté C.

A chaque individu ω peut être associée sa classe $C(\omega)$. C peut prendre ses valeurs dans ensemble quelconque pouvant être discret ou continu.

$$C : \Omega \mapsto \Gamma$$
$$\omega \to C(\omega)$$

Par exemple, si la population Ω est celle des clients d'une banque et C le résultat de la demande de crédit : *accepté* noté c_1 ; *refusé* noté c_2, alors $C(\omega)$ sera le résultat de la demande de crédit de l'individu ω. Dans le cas où C prend ses valeurs dans un ensemble discret, on parlera dans ce cas de classes.

Dans la réalité, l'observation de $C(\omega)$ n'est pas toujours facile pour diverses raisons. C'est pourquoi nous cherchons un moyen φ pour prédire la classe C et ainsi anticiper la décision de la banque.

La détermination du modèle de prédiction φ est liée à l'hypothèse selon laquelle les valeurs prises par la variable statistique C ne relèvent pas du hasard mais de certaines situations particulières que l'on peut caractériser à partir d'une série de données. Pour cela l'expert du domaine concerné établit une liste *a priori* de variables statistiques appelées « variables exogènes » et notées : $X = (X_1, X_2, \ldots, X_p)$. Ces variables sont également appelées « attributs prédictifs » dans le cas où C est continue et , « attributs explicatifs » dans le cas où C est discrète.

Les variables exogènes prennent leurs valeurs dans un espace de représentation noté $\Re$ qui peut avoir ou non une structure mathématique particulière: espace vectoriel par exemple.

$$X : \Omega \mapsto \Re$$
$$\omega \to X(\omega) = \left(X_1(\omega), X_2(\omega), \ldots, X_p(\omega)\right)$$

La variable X_1 pourrait être « l'âge », la variable X_2 « la situation de famille du demandeur », etc.

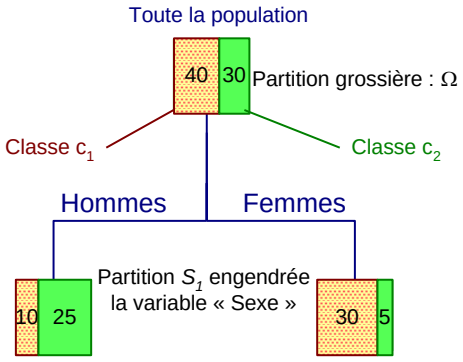
L'objectif est de rechercher un modèle de prédiction φ permettant, pour un individu ω issu de Ω, pour lequel nous ne connaissons pas $C(\omega)$ mais dont nous connaissons l'état de toutes ses variables exogènes $X(\omega)$ de prédire cette valeur grâce à φ.

- **Les graphes d'induction**

Les graphes d'induction, dont les modèles les plus populaires sont les arbres de décision, connaissent un grand succès : ils sont faciles à mettre en œuvre, les résultats qu'ils fournissent sont aisés à interpréter et les modèles qui en sont déduits sont performants. Les graphes d'induction sont capables d'appréhender des bases de données de grandes tailles et applicables sans restriction sur des données de n'importe quel type (qualitatives, quantitatives, ou un mélange des deux). La prolifération des logiciels fondés sur ce paradigme est d'ailleurs un signe assez fort de ce succès.

Toutes les méthodes à base de graphes d'induction fonctionnent sur le même principe que l'on peut décrire par l'algorithme général suivant :

On part de la partition grossière formée de tous les individus de l'échantillon d'apprentissage, on recherche ensuite, parmi les p variables exogènes $(X_1, X_2, \ldots, X_p)$, celle qui permet d'engendrer la meilleure partition au sens d'un critère donné. Celui-ci devra être d'autant meilleur que les classes de la partition sont homogènes. Nous obtenons un arbre à deux niveaux dont la racine représente la partition grossière Ω et dont les feuilles représentent les modalités de la variable exogène.



• Figure 19 : Arbre de décision

Sur la figure 19, nous supposons que population est composée de 70 individus dont 40 ont pris la valeur c_1 sur C, nous dirons qu'ils appartiennent à la classe c_1, et 30 appartiennent à la classe c_2. Les critères d'homogénéité d'une partition qui sont les plus fréquemment utilisés sont les mesures d'entropie. Si nous adoptons comme critère la mesure de l'entropie de Shannon dont l'expression générale est donnée par la formule suivante :

$$H(\Omega) = -\sum_{j=1}^{m} p(c_j/\Omega)\log_2 p(c_j/\Omega) \qquad (0.1)$$

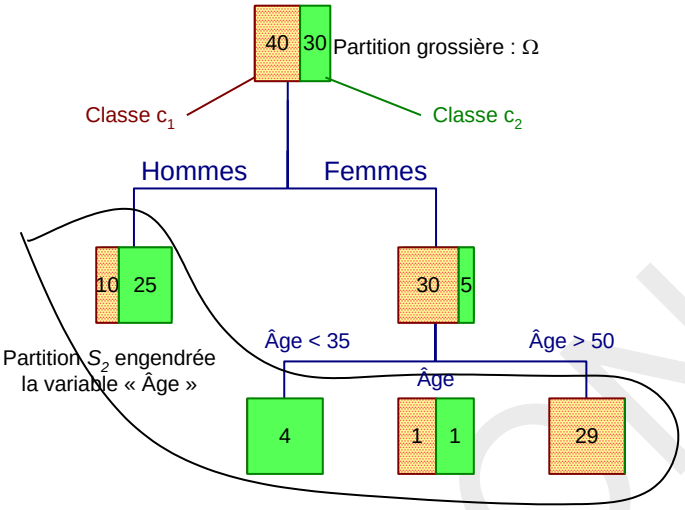
on aura $H(\Omega) = -\frac{40}{70}\log_2\frac{40}{70} - \frac{30}{70}\log_2\frac{30}{70} = 0{,}9852$. Sur la partition S_1 engendrée par la variable exogène « sexe », on calcule la valeur moyenne des entropies en chaque sommet :

$$H(S_1) = \sum_{s\in S_1} p(s)H(s) \;\; ; \qquad H(s) = -\sum_{j=1}^{m} p(c_j/s)\log_2 p(c_j/s) \qquad (0.2)$$

$$H(S_1) = \left[\left(\frac{35}{70}\right)\left(-\frac{10}{35}\log_2\frac{10}{35} - \frac{25}{35}\log_2\frac{25}{35}\right)\right] + \left[\left(\frac{35}{70}\right)\left(-\frac{30}{35}\log_2\frac{30}{35} - \frac{5}{35}\log_2\frac{5}{35}\right)\right] = 0{,}7273$$

La différence $H(\Omega) - H(S_1) = G_1$ est appelée le gain informationnel. C'est un moyen plus naturel pour mesure l'importance d'une variable. On cherchera la variable qui réduit l'entropie ou bien celle qui maximise le gain G.

Partant de la partition S_1, il s'agit de chercher une nouvelle partition S_2 qui soit meilleure que S_1 au sens du critère considéré.



• Figure 20 : arbre à deux niveaux

On peut vérifier que $H(S_1) - H(S_2) = G_2 > 0$. On continue ainsi jusqu'à ce qu'il ne soit plus possible de réduire la valeur du critère *H*.

Ce principe général a été décliné en plusieurs algorithmes. Ainsi en choisissant d'autres mesures de qualité, en imposant de n'engendrer que des bi-partitions, en autorisant les regroupements de modalités d'une même variable, en regroupant des modalités issues de pères différents pour obtenir des graphes latticiels, ou en introduisant des procédures d'élagage, on obtient une méthode particulière parmi la dizaine proposée dans la littérature. On trouve dans la référence Zighed et Rakotomalala une description détaillée des principaux algorithmes proposés.

Comme nous pouvons le constater sur le graphe de la figure 20, chaque branche représente une règle de la forme ***Si*** *Condition* ***alors*** *Conclusion*. La condition est une proposition en logique d'ordre 0 et la conclusion est la classe majoritaire sur le sommet. Ainsi, sur le graphe de la figure 20, on peut établir un modèle de décision φ composé de 4 règles :

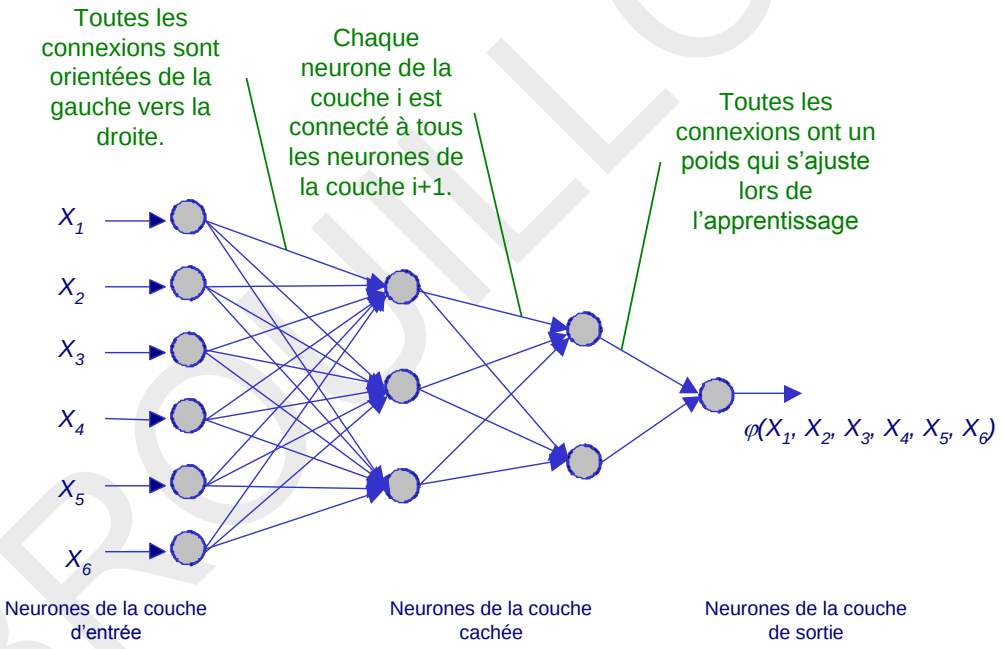
- **Si** « Sexe=Homme » **alors** $C = c_2$ (**risque d'erreur de**=10/35).
- **Si** « Sexe=Homme et Âge<35» **alors** $C = c_2$ (**risque d'erreur de**=0).
- **Si** « Sexe=Homme et Âge>50» **alors** $C = c_1$ (**risque d'erreur de**=0).
- **Si** « Sexe=Homme et 35≤Âge≤50» **alors** *C* = *indéterminé*

Grâce à ces règles, quand un nouvel individu se présente, connaissant son âge et son sexe, on peut prédire sa classe avec une certaine fiabilité.

Ces connaissances exprimées à travers ce modèle sont évaluées pour s'assurer de leur validité. Nous verrons, dans la prochaine section, les stratégies de validation qui sont en fait assez générales quel que soit le modèle.

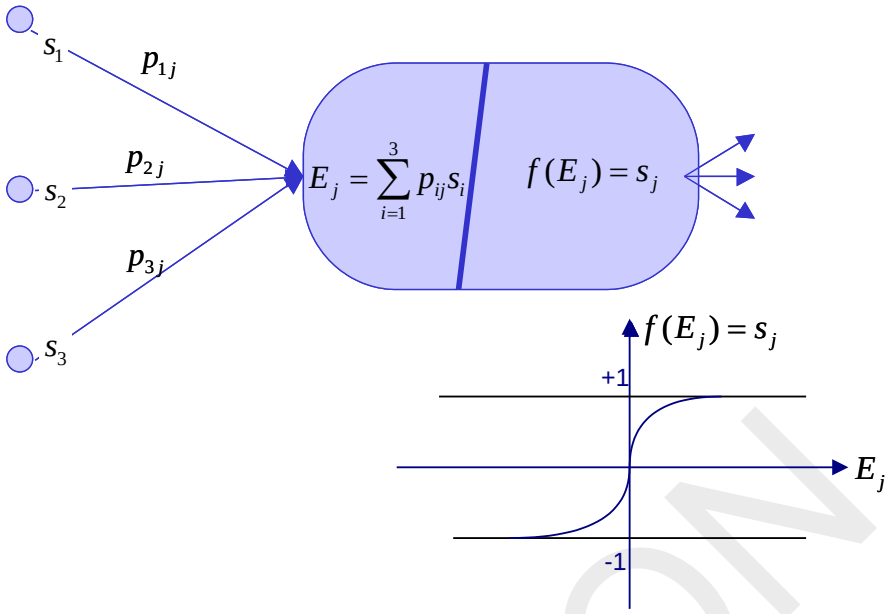
- **Les réseaux de neurones**

Les réseaux de neurones sont parmi les outils de modélisation les plus utilisés, en particulier pour les problèmes difficiles où le prédicteur que l'on cherche à construire repose sur de nombreuses interactions complexes entre les attributs exogènes. La structure générale d'un réseau de neurone se présente comme suit :



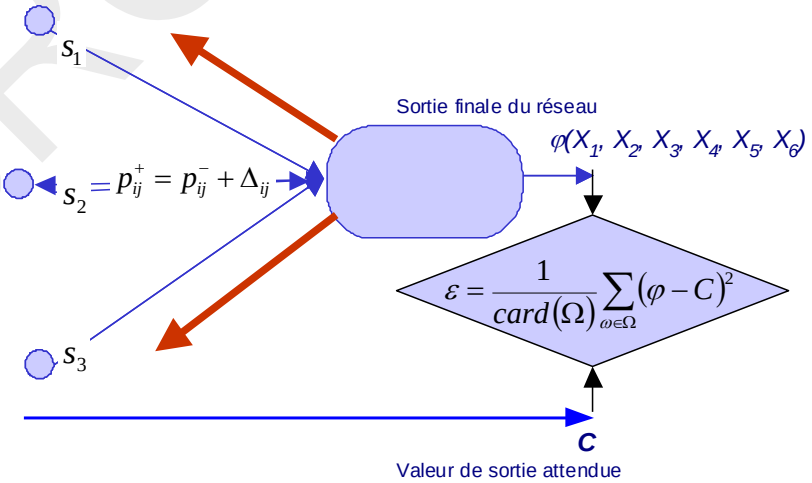
• Figure 21 : réseau multicouche

Les individus de l'échantillon d'apprentissage sont présentés avec leurs valeurs sur toutes les variables exogènes notées X_i. Sur cet exemple, les 6 valeurs sont transmises aux trois neurones de la première couche cachée. L'entrée du neurone de la couche cachée est en fait la somme des 6 valeurs d'entrée pondérées par les poids des connexions qui relient les deux couches. Chaque neurone de la couche cachée transforme son entrée (somme des entrées pondérées) en une sortie suivant un filtre qui est généralement une fonction à seuil comprise entre –1 et +1.



• Figure 22 : Propagation du signal d'une couche à la suivante

La valeur résultant du ou des neurones de la couche de sortie est comparée aux valeurs attendues correspondant à la variable (ou vecteur) endogène. Si l'écart est jugé « faible » sur un nombre suffisant d'itérations on considère alors que le réseau a appris le problème. L'apprentissage consiste à trouver les poids de toutes les connexions de sorte que l'erreur moyenne de prédiction, calculée sur les individus de la base d'apprentissage, soit jugée faible. Dans le cas où cette erreur est considérée comme importante, les poids de toutes les connexions sont modifiés de l'arrière vers l'avant selon une règle dite de rétropropagation de l'erreur, version dérivée de l'algorithme de descente de gradient.



• Figure 23 : Rétropropagation de l'erreur

Dans les réseaux de neurones, la principale difficulté réside dans le choix de la bonne architecture : nombre de couches cachées et nombre de neurones par

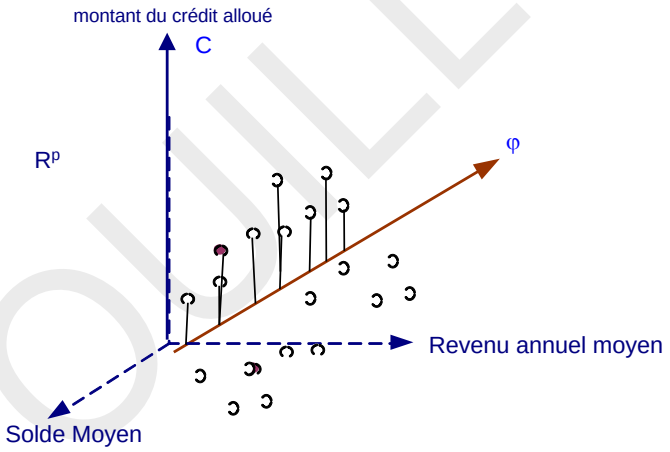
couche cachée. Les couches d'entrée et de sortie sont déterminées par la nature du problème : le nombre de neurones de la couche d'entrée est le plus souvent égal au nombre de variables exogènes et le nombre de neurones sur la couche de sortie est égal au nombre de variables endogènes.

La littérature sur les réseaux de neurone est très abondante. On y trouve de nombreux autres modèles.

- **Les méthodes de régression**

En régression, il s'agit d'expliciter une relation de type linéaire ou non entre un ensemble de variables exogènes et une variable endogène. Généralement, dans le cadre de la régression, toutes les variables sont considérées continues.

La régression linaire est dite multiple quand le nombre de variables exogènes est supérieur ou égal à deux.



• Figure 24 : Régression linaire, droite d'ajustement des moindres carrés

La relation que l'on cherche entre la variable endogène et les variables exogènes est linéaire et peut s'exprimer comme suit :

$$C = \alpha + XB \qquad (0.3)$$

où X est un vecteur ligne $X = (X_1, X_2, \ldots, X_p)$, B un vecteur colonne à p dimensions contenant les coefficients de la régression et α une erreur.

Nous allons chercher les coefficients de la droite φ qui ajuste au mieux le nuage de points au sens des moindres carrés. Nous pouvons trouver B en résolvant le problème d'optimisation suivant :

$$\min_{\Omega}\left(\sum_{\Omega}(C-\varphi)^2\right) \qquad (0.4)$$

où $\varphi = XB$.

La régression linéaire possède de nombreux résultats statistiques intéressants permettant d'apprécier la qualité du modèle qu'elle produit. La régression est probablement la technique de modélisation la plus utilisée.

La régression non linéaire vise à fournir un cadre dans lequel on puisse prendre en compte des liaisons polynomiales par exemple. Mais, dans la réalité, on s'arrange toujours pour ramener le problème au cas que l'on sait bien résoudre, c'est-à-dire linéaire.

En fait, la régression a été généralisée dans un cadre unique appelé « la régression linéaire généralisée » qui peut même être utilisée sur des variables catégorielles. La régression logistique permet de calculer la probabilité d'un événement endogène comme une combinaison linéaire de variables exogènes.

- **L'analyse discriminante**

L'analyse discriminante est l'une des plus anciennes techniques de discrimination. Elle a été proposée par Fischer en 1936. Cette technique est restée très populaire. Le problème pratique traité par Fischer pour illustrer cette méthode concerne la discrimination de trois classes de la famille des iris (versicolor, setosa et verginica). Pour cela, il a pris un échantillon de 150 iris répartis sur les trois classes et, pour chaque fleur, il a mesuré la longueur et la largeur des pétales et des sépales. Le fichier des ces données se trouve dans la plus part des ouvrages consacrés à cette méthode et sur les sites *web* des communautés d'apprentissage automatique comme celui de l'Université de Californie à Irvine ou celui de la communauté de *data mining* comme Kdnuggets.

Le principe de l'analyse discriminante est relativement simple. Considérons p variables exogènes toutes quantitatives $(X_1, X_2, \ldots, X_p)$, $p = 4$ dans le cas des iris de Fischer, et une variable endogène C qualitative qui prend ses valeurs dans un ensemble $\Gamma = \{c_1, c_2, \ldots, c_m\}$, avec $m = 3$ dans le cas des iris. Supposons que ces variables ont été centrées, c'est-à-dire que $\overline{X}_j = 0$; $j = 1, \ldots, p$.

Le problème que résout l'analyse discriminante consiste à construire une variable *Y*, combinaison linéaire de $(X_1, X_2, \ldots, X_p)$, telle que :

- Pour tout individu ω **de la classe** c_k on ait $Y(\omega)$ « peu différent » de $\overline{Y}_k$; $k = 1,...,m$ où $\overline{Y}_k$ désigne la moyenne de Y dans la classe c_k. Nous pouvons traduire cela en exigeant que la dispersion de *Y* soit minimale dans chaque classe. Nous cherchons donc à minimiser la variance de *Y* à l'intérieure de chaque classe, ce qui donne globalement, pour toutes les classes, le critère de la variance intra classe dont l'expression est :

$$V_{intra} = \frac{1}{card(\Omega)} \sum_{k=1}^{m} card(\Omega_k) \; V_k(Y) \qquad (0.5)$$

où $card(\Omega_k)$ représente l'effectif de la classe c_k et $V_k(Y)$ désigne la variance de Y dans la classe c_k.

- Pour tout individu ω' n'appartenant pas à la classe c_k on ait $Y(\omega) \neq Y(\omega')$. Cela représente le contraste entre les classes. Pour que ce contraste entre classes soit le plus fort, on cherchera à déterminer Y telle que les moyennes $\overline{Y}_k$; $k = 1,...,m$ soient les plus dispersés possibles. Cela revient à maximiser la variance interclasses dont l'expression est :

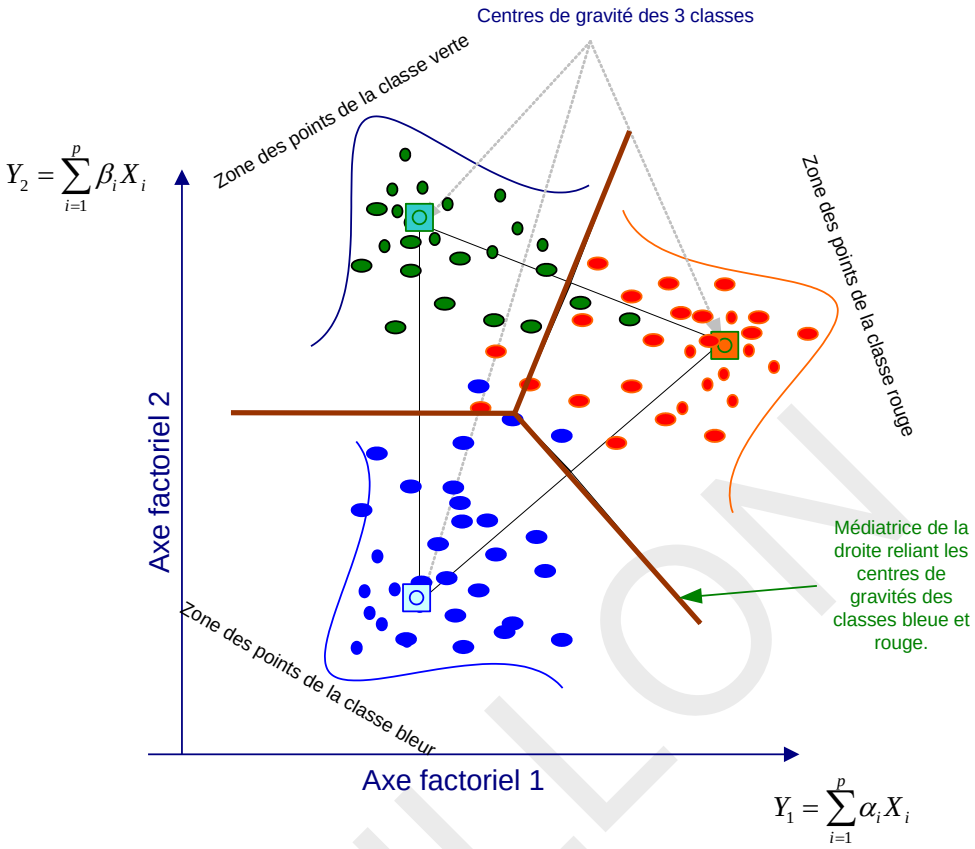
$$V_{inter} = \frac{1}{card(\Omega)} \sum_{k=1}^{m} card(\Omega_k) \; \left(\overline{Y}_k - \overline{Y}\right)^2 \qquad (0.6)$$

Nous sommes ainsi en face d'un problème d'optimisation que l'on formule ainsi :

Trouver Y, une combinaison linéaire des variables exogènes $(X_1, X_2, \ldots, X_p)$, qui minimise la variance intra classes et qui maximise la variance inter classes. Ce problème se résout facilement car cela revient à chercher des valeurs propres et des vecteurs propres associés à une matrice de variances.

Dans le cas où nous chercherions à discriminer entre m classes, on démontre aisément que l'on peut trouver au plus $(m-1)$ droites. Ces droites s'appellent axes factoriels discriminants. On peut donc les utiliser par paire pour visualiser les classes.

Un nouvel individu à classer est affecté à la classe dont le centre de gravité est le plus proche. On peut définir géométriquement des surfaces de décision par l'intersection des médiatrices sur les droites qui relient les centres de gravité des classes comme illustré par la figure suivante :



• Figure 25 : Plans factoriels discriminants

La présentation de l'analyse discriminante que nous venons de donner repose sur trois hypothèses : toutes les variables sont quantitatives, elles sont normalement distribuées et les classes sont linéairement séparables. Ces hypothèses sont hélas rarement toutes vérifiées. De nouveaux développements ont été introduits pour élargir le champ d'application à des frontières quadratiques ou à des données catégorielles.

- **Les réseaux bayésiens**

Les réseaux bayésiens sont apparus au début des années 1980. Rendus populaires par le groupe de recherche de la firme Microsoft qui les introduits dans les systèmes d'aide contextuelle d'Office, ils sont maintenant très utilisés dans la modélisation des processus complexes de décision.

L'idée de base des réseaux bayésiens repose sur le célèbre théorème de Bayes mais s'est considérablement enrichie ces vingt dernières années.

Considérons un échantillon Ω sur lequel nous avons observé p variables statistiques $(X_1, X_2, \ldots, X_p)$. Cette observation se matérialise par un tableau de données D de p colonnes et n lignes si n est la taille de l'échantillon.

Soit H une hypothèse qui pourrait être : les individus de Ω caractérisés par D appartiennent à la classe c_k. Nous cherchons à évaluer cette hypothèse H, connaissant D, ce que nous pouvons traduire par la détermination $P(H/D)$ et qui n'est autre que sa probabilité *a posteriori*.

Prenons un exemple dans lequel tous les individus de Ω sont décrits par deux variables booléennes qui nous renseignent si la personne fume ou pas et si elle a ou non dans sa famille des cas de cancer. On note F la variable « fumeur » et A « antécédents familiaux » qui peuvent être vrai ou faux. On s'intéresse au fait qu'une personne a ou n'a pas un cancer que l'on note C. Prenons un individu particulier dont on sait qu'il fume et qui n'a pas d'antécédents familiaux. Dans ce cas, notre tableau de données D comporte une seule ligne : F = vrai, A = faux. Il s'agit alors de déterminer la probabilité $P(C/D)$ que cet individu a le cancer sachant ses données D.

On note $P(C)$ la probabilité *a priori* qu'a une personne d'être atteinte d'un cancer, $P(D/C)$ la probabilité *a posteriori* d'observer une personne qui fume et qui n'a pas d'antécédents familiaux chez les personnes ayant un cancer et $P(D)$ la probabilité *a priori* d'être fumeur et de ne pas avoir d'antécédents familiaux.

La question qui se pose est de savoir comment estimer toutes ces probabilités. $P(C)$, $P(D)$ et $P(D/C)$ peuvent être estimées à partir d'un échantillon d'apprentissage et $P(C/D)$ en utilisant le théorème de Bayes dont l'expression est :

$$P(C/D) = \frac{P(D/C)P(C)}{P(D)} \qquad (0.7)$$

L'un des premiers modèles proposés est appelé le « bayésien naïf ». Il repose sur l'hypothèse selon laquelle les variables sont indépendantes, c'est-à-dire que le fait de fumer ou pas n'a aucun lien avec le fait d'avoir des antécédents familiaux, et réciproquement, ce qui se traduit par $P(F,A) = P(F).P(A)$. Comme C peut prendre deux états seulement « c_1 = cancer » et « c_2 = pas de cancer », on cherchera la conclusion donnée par la probabilité a posteriori maximale, c'est-à-dire :

$$\max_{k=1}^{2}\left[P(C = c_k/D) = \frac{P(D/C = c_k)P(C = c_k)}{P(D)} \right] \qquad (0.8)$$

Or, dans cette équation, la probabilité $P(D)$ est constante quelle que soit la classe c_k, on cherchera donc à ne maximiser que le numérateur,

$[P(D/C=c_k)P(C=c_k)]$. Puisque nous supposons l'indépendance des symptômes, on établit alors :

$$P(D/C=c_k)P(C=c_k) = P(F=vrai, A=\ faux/C=c_k)P(C=c_k)$$
$$= P(F=vrai/C=c_k)P(A=\ faux/C=c_k)P(C=c_k)$$

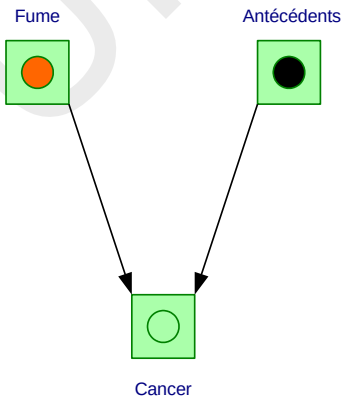
l'estimation de ces probabilités est faite par les fréquences observées dans l'échantillon. Cela montre qu'il faut disposer d'un échantillon suffisant pour que ces probabilités aient un sens.

Si $P(C=c_1/D) > P(C=c_2/D)$, on conclut alors que l'hypothèse la plus probable est la présence d'un cancer.

Il est clair que l'hypothèse d'indépendance ne peut pas être retenue dans la plus part des cas. C'est la raison pour laquelle ont été introduits les réseaux bayésiens.

Un réseau bayésien est représenté par un graphe acyclique dans lequel les sommets sont des variables booléennes et les arcs les relations de dépendance. L'architecture du réseau est généralement fournie par l'expert.

Considérons le réseau simple de la figure suivante :



• Figure 26 : Structure simple d'un réseau bayésien

Ce schéma illustre que la variable cancer est déterminée par deux facteurs que sont « Fume » et « Antécédent ». Ces deux facteurs sont soit vrai soit faux indépendamment l'un de l'autre.

On détermine ensuite, soit en interrogeant l'expert, soit par apprentissage à partir d'un échantillon, les probabilités des états de chaque nœud. Supposons ces valeurs déterminées.

$P(F = vrai) = 0.1;\ \ P(F = \ faux) = 0.9$
$P(A = vrai) = 0.25; P(A = \ faux) = 0.75$
$P(C = vrai / F = vrai, A = vrai) = 1; P(C = \ faux / F = vrai, A = vrai) = 0$
$P(C = vrai / F = \ faux, A = vrai) = 1; P(C = \ faux / F = \ faux, A = vrai) = 0$
$P(C = vrai / F = vrai, A = \ faux) = 0.9; P(C = \ faux / F = vrai, A = \ faux) = 0.1$
$P(C = vrai / F = \ faux, A = \ faux) = 0.05; P(C = \ faux / F = \ faux, A = \ faux) = 0.95$

A partir de ces informations, il est possible de calculer n'importe quel état parmi les 8 états possibles du système. Par exemple, supposons qu'un individu a un cancer et qu'il a des antécédents familiaux, on souhaite savoir quelle est la probabilité pour qu'il soit fumeur : $P(F = vrai / C = vrai, A = vrai)$

$$P(F = vrai / C = vrai, A = vrai) = \frac{P(F = vrai, C = vrai, A = vrai)}{P(C = vrai, A = vrai)}$$
$$= \frac{P(F = vrai, C = vrai, A = vrai)}{P(F = vrai, C = vrai, A = vrai) + P(F = \ faux, C = vrai, A = vrai)}$$

Or,
$$P(F = vrai, C = vrai, A = vrai) = P(C = vrai / F = vrai, A = vrai)P(F = vrai)P(A = vrai)$$

et

$$P(F = \ faux, C = vrai, A = vrai) = P(C = vrai / F = \ faux, A = vrai)P(F = \ faux)P(A = vrai)$$

Ce qui donne, d'après les valeurs numériques fournies plus haut :

$$P(F = vrai / C = vrai, A = vrai) = \frac{1 \times 0.1 \times 0.25}{1 \times 0.1 \times 0.25 + 1 \times 0.9 \times 0.25} = 0.1$$

Dans le cas de réseaux plus complexes, comportant plusieurs sommets et de nombreux arcs, ce calcul devient parfois difficile, voire impossible, en un temps raisonnable. Différentes stratégies ont été proposées pour réaliser ces calculs, nous renvoyons le lecteur aux articles de référence pour un approfondissement.

Dans un réseau bayésien comme celui-ci, on peut introduire des éléments de décision et de coût. Par exemple, on peut introduire le fait de traiter le patient selon le protocole de soin A ou B. A chacun des deux protocoles est associé un coût et on peut alors se demander ce qu'il faut faire suivant un état du système.

- **Les règles d'association**

La recherche de règles d'association dans une base de données est probablement le problème qui a le plus fortement contribué à l'émergence du *data mining* en tant que domaine scientifique à part entière. La grande distribution, les télécommunications et plein d'autres secteurs de la grande consommation enregistrent, dans un but de facturation, l'ensemble des transactions commerciales avec leurs clients. Pour une grande surface de distribution, cela peut atteindre plusieurs centaines de millions de transactions effectuées par jour. Les données enregistrées au passage en caisse servent d'une part à la facturation au client et d'autre part à des actions de gestion comme le suivi des stocks ou encore à l'étude de la composition des paniers dans un but de marketing. Par exemple, si dans un nombre de cas significativement grand on a observé que les clients qui achètent les couches culottes achètent également un pack de bières, alors il peut être judicieux de disposer ces deux articles côte à côte dans le magasin si le but est de permettre au client de trouver vite les produits qu'il cherche. L'étude des transactions en vue d'identifier des associations entre produits permet de mieux caractériser un client et ainsi de définir des actions commerciales ciblées envers lui. Ainsi, si vous visitez le site de vente de livres en ligne amazon.com et que vous recherchez des livres sur le *data mining*, vous pouvez voir ressortir des livres de statistique mais pas de livres de mathématiques. Ce résultat peut être issu du fait que l'étude des transactions des anciens clients a montré qu'un nombre significatif de ceux qui ont acheté un ou plusieurs livres sur le *data mining* ont également acheté un ou plusieurs livres de statistique mais que l'inverse n'est pas nécessairement vrai.

Les p variables statistiques $\mathrm{X} = (X_1, X_2, \ldots, X_p)$ sont toutes booléennes et sont appelées « items ». La population Ω est composée de l'ensemble des transactions. Comme toutes les variables sont booléennes, une transaction ω est caractérisée par un sous-ensemble de variables $\mathrm{X}' \subseteq \mathrm{X}$ (ou ensemble d'items) toutes à 1.

Soit X_a un ensemble d'items, $\mathrm{X}_a \subseteq \mathrm{X}$; une transaction ω contient l'ensemble d'items X_a si et seulement si tous les items qui forment X_a sont à 1 dans la transaction ω. Par exemple, si X_a désigne quatre produits du magasin que l'on note A, B, C et D, on dira que X_a est contenu dans la

transaction ω si les quatre produits figurent dans la transaction, on écrira alors $\mathrm{X}_a \subseteq \omega$.

Une règle d'association est une implication de la forme $\mathrm{X}_a \Rightarrow \mathrm{X}_b$ où $\mathrm{X}_a \subset \mathrm{X}$ et $\mathrm{X}_b \subset \mathrm{X}$ et $\mathrm{X}_a \cap \mathrm{X}_b = \phi$. La règle $\mathrm{X}_a \Rightarrow \mathrm{X}_b$ se produit dans l'ensemble des transactions Ω avec un support s qui est le pourcentage de transactions de Ω qui contient $\mathrm{X}_a \cup \mathrm{X}_b$ et une confiance c qui est le pourcentage de transactions Ω contenant X_a qui contiennent également X_b. Ainsi,

$$support(\mathrm{X}_a \Rightarrow X_b) = P(\mathrm{X}_a \cup X_b) = s \text{ et}$$
$$confiance(\mathrm{X}_a \Rightarrow X_b) = P(\mathrm{X}_b / X_a) = c$$

Les règles de la forme $\mathrm{X}_a \Rightarrow \mathrm{X}_b$ sont dites pertinentes si $support(\mathrm{X}_a \Rightarrow X_b) > s_0$ et $confiance(\mathrm{X}_a \Rightarrow X_b) > c_0$ où s_0 et c_0 sont des seuils minimums.

La recherche de règles d'associations dans un ensemble de transaction s'opère en deux temps :

a) On cherche les ensembles d'items fréquents, c'est-à-dire ceux qui apparaissent un nombre minimum de fois dans l'ensemble des transactions.

b) On génère les règles d'associations pertinentes, c'est-à-dire celles qui vérifient simultanément la contrainte minimale sur le support et la confiance.

La recherche de tous les sous-ensembles fréquents consiste à déterminer parmi l'ensemble de toutes les parties de $\mathrm{X} = (X_1, X_2, \ldots, X_p)$ les sous-ensembles fréquents, c'est-à-dire présents dans un nombre assez conséquent de transactions. Il est facile de constater qu'une recherche exhaustive devient quasiment impossible à traiter dès lors que l'on dépasse quelques dizaines d'items car le nombre de parties d'un ensemble croît de façon exponentielle avec son cardinal.

L'algorithme de base Apriori propose une stratégie moins brutale. Elle préconise la recherche des ensembles fréquents de cardinal $k+1$ à partir des ensembles fréquents de cardinal k. Ainsi pour trouver les ensembles fréquents ayant deux items, on utilisera exclusivement les ensembles fréquents ayant un item. Il est clair qu'en toute généralité, le nombre d'ensembles fréquents diminue avec le nombre d'items : il y a moins d'ensemble fréquents à 2 items qu'à un item. Cette propriété permet ainsi de

restreindre la taille de l'espace à explorer pour trouver tous les ensembles fréquents nécessaires à la deuxième étape de l'algorithme qui comporte deux points :

i) Pour chaque ensemble fréquent X_a on génère tous les sous-ensembles non vides.

ii) Pour chaque sous-ensemble non vide $\mathrm{X}_b \subset \mathrm{X}_a$, on génère la règle $(\mathrm{X}_b \Rightarrow (X_a - X_b))$ si $\frac{support(\mathrm{X}_a)}{support(\mathrm{X}_b)} > c_0$

Plusieurs variantes de l'algorithme Apriori ont été proposées pour améliorer ses performances.

- **Autres approches en prédiction**

Il existe encore de nombreuses autres techniques destinées à la prédiction parmi lesquelles on peut citer CN2 qui opère sur des données discrètes et qui génère des règles logiques de la forme « si condition alors conclusion », les SVM (*support vector machine*) issus des travaux sur la théorie de l'apprentissage ou les méthodes à base de voisinage comme les « k plus proches voisins » qui, pour un point dont on souhaite déterminer la classe d'appartenance, utilisent l'information sur les classes dans le voisinage du point en question. Nous pouvons également citer les algorithmes génétiques pour modéliser un apprentissage de règles, ou le raisonnement à partir de cas. Toutes ces méthodes sont décrites dans de nombreux ouvrages cités en bibliographie.

Si on dénombre les différentes variantes des algorithmes de prédiction, il en existe plusieurs dizaines et doit s'en créer de nouvelles méthodes régulièrement. A ce jour, aucune méthode ne s'est avérée supérieure en tous points aux autres. Il y a celles qui sont mieux adaptées aux petits échantillons, celles qui ne travaillent que sur des données numériques, celles qui fonctionnent en boîte noire, celles qui ne fournissent pas une fonction de classement explicite, celles qui ne savent travailler que sur deux classes, etc.

Le choix d'une méthode parmi celles-ci s'avère donc difficile car il dépend de plusieurs facteurs : taille de l'échantillon, nature des variables exogènes (qualitative, quantitatives ou mixtes), la nature de la variable endogène, l'intelligibilité du modèle, la connaissance *a priori* que nous avons sur la structure des classes (linéairement séparables ou pas), la complexité de l'algorithme, etc.

10. Phase de validation et de mise en forme

Les modèles extraits, notamment par les méthodes d'apprentissage supervisé, ne peuvent être utilisés directement en toute fiabilité. Nous devons les évaluer, c'est-à-dire les soumettre à l'épreuve de la réalité et apprécier leur justesse. Le procédé habituel consiste à estimer au mieux le taux d'erreur du modèle. Ainsi, l'utilisateur décidera d'appliquer ou non le modèle de prédiction en connaissance des risques qu'il prend. Le taux d'erreur est généralement calculé à partir de la matrice de confusion. Celle-ci donne le pourcentage d'affectation dans les différentes classes en fonction des classes d'origine.

Matrice de confusion. En ligne les classes d'origine et en colonnes celles d'affectation		**Classes d'affectation**		
		$\hat{c}_1$	.....	$\hat{c}_n$
Classes d'origine	c_1	n_{11}		n_{1m}
	:	:	:	:
	:	:	:	:
	c_n	n_{m1}	.....	n_{mm}

• Tableau 1 : matrice de confusion

L'effectif n_{ij} désigne le nombre de cas de la classe c_i affectés à tord à la classe c_j . Ainsi

$$S = \sum_{i=1}^{m} n_{ii} \tag{0.9}$$

représente le nombre d'affectations correctes. Le taux d'erreur est alors calculé

$$E = 1 - \frac{\sum_{i=1}^{m} n_{ii}}{\sum_{\substack{i=1 \\ j=1}}^{m} n_{ij}} \tag{0.10}$$

Ce taux d'erreur calculé sur l'échantillon d'apprentissage est généralement optimiste, c'est-à-dire plus faible que le vrai taux d'erreur inconnu, celui qu'on aurait eu si on avait l'exhaustivité de la population concernée. On appelle ce taux d'erreur « taux d'erreur en resubstitution ». Généralement on lui préfère le taux d'erreur en validation calculé de la même façon mais sur un nouvel échantillon dit de validation qui n'a pas servi lors de la phase d'apprentissage.

Pour approcher au mieux la valeur du taux d'erreur théorique, on utilise différentes techniques de rééchantillonnage.

- **La validation croisée**

La validation croisée, ou *cross validation*, consiste à répartir l'échantillon d'apprentissage aléatoirement en K paquets d'effectifs identiques. Si on note Ω_k, $k = 1,..., K$ les différents sous-échantillons, le taux d'erreur en validation croisée est calculé en réservant, à tour de rôle, un échantillon Ω_k qui servira à mesurer le taux d'erreur en validation, l'apprentissage étant réalisé sur la totalité des individus restants : $\Omega - \Omega_k$. On obtient ainsi K taux d'erreurs E_k, $k = 1,..., K$. Le taux d'erreur en validation croisée est alors la moyenne arithmétique des taux d'erreurs partiels :

$$E = \sum_{k=1}^{K} E_k \qquad (0.11)$$

Généralement on lui associe la variance $\sigma_E^2 = \frac{1}{K}\sum_{k=1}^{K}(E_k - E)^2$.

Dans le cas où le découpage de l'échantillon conduit à avoir un individu par paquet, c'est-à-dire $K = n$ la méthode est connue sous le nom « *leave one out* ».

- **Le *bootstrap***

Il s'agit d'une méthode dans laquelle l'échantillon de validation est tiré aléatoirement à chaque itération. Contrairement à la validation croisée, un même individu statistique peut être pris dans plusieurs échantillon, ou jamais. Les paramètres fournis à l'utilisateur sont le nombre de répétitions du tirage et la taille de l'échantillon de validation. De façon analogue à la validation croisée, l'erreur du modèle est estimée par l'erreur moyenne réalisée sur les différents échantillons de validation.

Les méthodes de validation dont nous venons de parler fonctionnent toutes *a posteriori*, après apprentissage, et elles sont globales à chaque modèle.

La mise en forme des modèles comporte différents aspects, allant de la visualisation des connaissances pour les rendre intelligibles jusqu'à l'agrégation des modèles, en passant par leur simplification.

11. Le *text mining*

Les données textuelles en format « libre » disponibles sur support informatique représentent 70% des données numériques et se retrouvent sous forme de rapports, de courriers, de publications, de manuels, etc. Les textes contiennent des informations et des connaissances utiles et parfois critiques pour la gestion et la prise de décision dans les entreprises.

La fouille de données textuelles, ou *text mining*, vise à définir des stratégies pour exploiter les textes en format libre. On y distingue deux niveaux de traitement.

- Le premier niveau porte sur la recherche d'information dans les bases de données textuelles. On y trouve essentiellement des outils de requête en langage naturel. Par exemple, rechercher les textes qui contiennent les mots X et Y ou Z. Grâce au développement des technologies du traitement de la langue naturelle, on peut également formuler des requêtes plus complexes contenant des expressions ou même des textes en exemple. Les techniques de recherche d'information disponibles permettent, l'accès aux textes par les contenus à la fois lexicaux et sémantiques.

- Le second niveau porte plus spécifiquement sur l'extraction de connaissances à partir de données textuelles. En effet, certaines recherches en texte intégral peuvent être extrêmement difficiles car l'utilisateur ne sait pas toujours formuler une requête. Par exemple, comment rechercher, dans la presse électronique les textes qui incitent à la violence dans les stades. Bien évidemment l'utilisateur ne souhaite pas avoir la totalité des textes qui traitent de la violence tout court car il peut y en avoir des milliers ni ceux qui traitent de la violence sur les stades qui peuvent également être très nombreux, mais seulement ceux qu'un jugement humain peut considérer de nature à inciter les lecteurs à être violents sur les stades. De plus, ces textes peuvent ne contenir aucun des mots « violence » ou « stade ». C'est là qu'interviennent les méthodes de *data mining* qui peuvent aider l'usager à déterminer les règles minimales qui permettent de reconnaître ces textes. Ces règles minimales, une fois validées, deviennent alors des connaissances discriminatoires pour ce type de textes.

Il existe actuellement deux courants dans le domaine du *text mining*, le premier plutôt centré sur l'exploitation des textes dans le contexte de la langue (connaissances lexicales, grammaticales ou linguistiques), le second qui considère les textes en termes d'occurrences d'unités lexicales. Dans cette dernière approche, les unités lexicales sont définies par une fenêtre qui se déplace linéairement sur le texte. Elles peuvent être les mots de la langue rencontrés dans les textes analysés ou simplement des groupes de N lettres appelés des N-grammes. Par ce procédé, un corpus de Z textes est transformé en un tableau d'occurrences de Z lignes et K colonnes. Les lignes représentent les textes et les colonnes les unités lexicales présentes dans la totalité du corpus. A l'intersection de la ligne i et la colonne j figure α_{ij} ; $\alpha_{ij} \geq 0$ qui est généralement la fréquence relative du terme k_j dans le texte t_i. A partir de ce moment, le corpus de texte est regardé comme un tableau de données susceptible d'être traité par la totalité des méthodes de *data mining* évoquées plus haut. On peut ainsi chercher des fonctions discriminantes entre des classes de textes ou bien catégoriser au moyen des méthodes de classification les textes ou simplement de décrire les textes les uns par rapport aux autres en utilisant des méthodes comme l'analyse factorielle des correspondances.

Généralement, une opération de *text mining* suit plusieurs étapes :

1. La préparation des textes. Cette étape, qui consiste à nettoyer et enrichir les textes, se comporte plusieurs phases :

   - nettoyage des textes par la suppression des unités lexicales non informatives. Par exemple, l'usager peut considérer que les articles « un », « une », « le », « la », « les », « de » ou « des » ainsi que les signes de ponctuation ou les espaces en trop ne portent pas de sens et il les supprime de tous les textes ;

   - lemmatisation où toutes les déclinaisons d'un verbe ou d'un mot sont ramenées à la racine. Par exemple, les verbes conjugués sont ramenés à leur forme à l'infinitif ;

   - traitement des synonymes pour conduire l'usager à ne prendre qu'un seul terme ;

   - gestion des expressions ou mots composés, etc.

   On peut ainsi imaginer une série d'opérations de préparation dans le cadre de langue utilisée.

2. Le choix du type de codage (par N-grammes, mots, syntagmes nominaux ou par un mélange de ces derniers). A ce stade, on peut obtenir des milliers d'unités qui deviennent des attributs dans le tableau des données.

3. La mise en œuvre des méthodes de sélection des attributs pour en réduire le nombre. Les attributs créés sont généralement très grand, de plusieurs centaines et jusqu'à plusieurs milliers. Cette réduction se fait soit par sélection d'attributs redondants (très fortement corrélés entre eux), soit par changement d'espace de description (en travaillant dans l'espace des premiers facteurs principaux issus d'une analyse factorielle des correspondances).

4. La mise en œuvre de techniques de *data mining* sur les données obtenues après filtrage et transformations des étapes antérieures. A ce stade, toutes les méthodes de *data mining* décrites peuvent être utilisées selon la nature du problème posé.

11. L'*image mining*

A l'instar des données textuelles, les données sous formes d'images peuvent également être traitées par les techniques de *data mining* en vue d'extraire des connaissances. Celles-ci permettraient d'identifier, de reconnaître ou de classer automatiquement des bases volumineuses d'images. Actuellement, les principales techniques d'interrogation dans des bases d'images utilisent des fichiers d'index. A chaque image est associée une série d'index qui donnent des indications sur son contenu. Le plus souvent, cette indexation est effectuée manuellement. Les techniques de *data mining* sont de plus en plus utilisées pour automatiser ces opérations.

Pour être exploitées par des méthodes de *data mining*, les images doivent également subir une série de pré-traitement en vue d'obtenir des tableaux numériques. Les principales étapes du pré-traitement sont les suivantes :

1. Transformation, filtrage et mise en forme. Les usagers sont souvent conduits à modifier les images initiales afin de mieux faire ressortir certaines caractéristiques qu'ils considèrent comme importantes. Par exemple, décider d'accentuer le contraste sur l'ensemble des images par des transformations mathématiques appropriées.

2. Extraction de caractéristiques. Pour être traitées par des techniques de *data mining*, les images doivent être transformées en un ensemble de vecteurs de nombres. Ensuite, la banque d'images est représentée par un tableau numérique. Chaque ligne étant une image et chaque colonne une caractéristique sur l'image. Cette opération de vectorisation peut se faire de deux façons :

   - la première consiste à mettre bout à bout toutes les lignes de la grille de l'image. Ainsi, pour une image à niveaux de gris de taille 512 x 512, nous obtenons un vecteur de 512 x 512 = 262 144 colonnes ayant chacune une valeur comprise entre 0 et 255 (correspondant au niveau de gris du pixel considéré). S'il s'agit d'une image en couleur, ce nombre de colonnes sera multiplié pour chacune des trois couleurs de base (RVB) ;

   - la seconde façon consiste à calculer une série de caractéristiques globales ou locales sur chaque image. Par exemple, une image peut être codée par 256 attributs qui décrivent l'histogramme des niveaux de gris, 13 attributs qui décrivent la texture, etc. On obtient ainsi K attributs numériques pour chaque image. Généralement, les caractéristiques globales sont assez peu informatives car l'information locale joue un rôle important, notamment dans l'identification des objets présents sur l'image. Pour cette raison, on procède à un codage différent. L'image est alors découpée en K pavés de taille $x \times y$ pixels qui peuvent être recouvrants. Les attributs sont alors calculés sur chaque pavé de l'image.

3. Mise en œuvre des méthodes de *data mining*. A l'issue de l'étape précédente, le corpus d'image est transformé en un tableau de données numériques sur lesquelles nous pouvons mettre en œuvre les techniques d'analyse, de classification ou d'apprentissage.

13. Le *multimedia mining*

Le *multimedia mining* obéit aux mêmes principes que ceux que nous avons établis pour le texte ou les images, à savoir la définition des transformations, des filtres et le recodage de la séquence vidéo en tableaux numériques.

Dans le multimédia, nous avons au moins deux objets à coder : les images et le son. Mais nous pouvons disposer également de données textuelles, des images des séquences vidéo ou même d'hypertextes. Cela ne change en rien aux problèmes à résoudre et qui sont essentiellement ceux du codage de l'information brute.

14. Le *web mining*

Les réseaux électroniques, de l'intranet à Internet, constituent une formidable source d'informations de par son large volume (accroissement exponentiel), sa richesse (données multimédias) et sa densité (souvent plusieurs milliers de pages pour une thématique). Les intérêts de fouiller dans ces données sont multiples et variés mais se heurtent finalement à deux problématiques majeures ; l'une concernant « l'internaute » et l'autre concernant « le diffuseur » des informations.

La problématique de l'internaute peut se résumer à celle de la recherche et l'analyse d'informations pertinentes, par exemple dans le domaine de la veille technologique. En s'appuyant sur les données présentes sur ces réseaux (des pages html, des liens Internet, des fichiers de log), les techniques de *data mining*, de *text mining* et d'*image mining*, peuvent offrir des solutions intéressantes. Cette problématique rejoint d'une certaine façon celle du *multimedia mining*. La seule différence est que, dans ce cas, la source d'information est le *web*. L'usager définit une stratégie pour rapatrier des informations depuis le *web*. Il peut indiquer la profondeur de la recherche, le type d'objets à rapatrier, les sites à exclure, etc. Il dispose alors d'un corpus de données hétérogènes qu'il doit coder sous forme d'un tableau numérique pour ensuite lancer des méthodes de *data mining* sur ces données et en extraire des connaissances.

La problématique du diffuseur ou du propriétaire de sites *web* consiste à déterminer les différents profils d'internautes en fonction de leur parcours sur le site afin de pouvoir cibler ses offres, orienter son discours et donc de proposer rapidement les informations recherchées aux clients potentiels.

Les propriétaires de sites Internet sont quant à eux intéresser par les visiteurs. A chaque passage sur les pages *web*, un internaute laisse des traces sur les sites visités. Outre la date et l'heure de la visite, le site hôte enregistre le numéro de la machine, le navigateur utilisé, l'ensemble des pages visitées, etc. L'exploitation et la fouille de ces données constitue une source d'information intéressante pour le propriétaire du site.

15. Les grandes applications

15.1. La gestion de relation client

Dans un contexte concurrentiel de plus en plus soutenu, la capacité à conquérir et à retenir les clients repose sur une connaissance fine de leurs besoins et de leur comportement de consommateurs, d'utilisateurs et d'acheteurs.

Créer et maintenir une relation de plus en plus personnalisée, à partir d'un produit de plus en plus standardisé, est un facteur clé de succès pour les produits et services de grande consommation.

Pour cela, il faut bien connaître ses clients. Les études de marché réalisées selon les méthodes classiques sont longues à mettre en œuvre, pour des résultats toujours sujets à caution en raison de la taille des échantillons pratiqués. L'entreprise dispose pourtant d'informations sur ses clients et leurs habitudes de consommation.

Au sein de la gestion de relation client, ou *customer relationship management* (CRM), on peut distinguer trois dimensions :

Le CRM opérationnel concerne gestion des relations avec les clients, qui constituent le *front office* du dispositif. Le CRM opérationnel permet d'emmagasiner des informations permettant une connaissance approfondie des clients, et dont le stockage est désormais rendu possible par les technologies de *data warehouse*.

Le CRM analytique consiste en l'exploitation des bases de données créées par l'entreprise sur ses clients. C'est à ce stade qu'entrent en jeu les techniques de *data mining*.

Le CRM collaboratif vise à intégrer des outils communiquant dans les dispositifs de front office, afin d'optimiser les échanges d'informations dans la gestion quotidienne des activités commerciales.

Toutes les entreprises, notamment de la grande distribution où la concurrence est forte, mettent en place des structures pour le CRM analytique. Les objectifs des analyses en *data mining* sont multiples : segmentation de la clientèle, fidélisation de la clientèle, organisation de rayons de magasins, etc.

15.2. L'aide à la décision dans les processus industriels

L'automatisation est l'un des meilleurs facteurs d'augmentation de la productivité. Les industriels, notamment dans la surveillance, le diagnostic et la maintenance des unités de production ont depuis longtemps fait appel aux méthodes de la statistique et de la modélisation. Il est donc naturel pour cette activité d'incorporer le *data mining* pour mieux analyser les pannes ou organiser les ateliers.

15.3. La génomique

Toute espèce vivante, animale ou végétale est composée de cellules. L'homme en possède quelques 100 000 000 000 000 000 000 (cent mille millions de milliard). Chaque cellule contient la totalité du génome, c'est-à-dire tout ce qui forme l'identité biologique de l'individu, son patrimoine génétique. Toutes les cellules d'un même individu contiennent ainsi la même copie du génome. Le génome de toutes les espèces est formé d'une

succession de quatre bases chimiques désignées par les lettres A, C, T et G. L'alignement de ces bases, qui forment la molécule d'ADN, peut être vu comme un texte écrit dans un alphabet de 4 lettres (A,C,T et G). Le nombre de lettres présentes dans le génome humain est de 3 milliards environs. Si on devait transcrire le code génétique d'une personne sur du papier en y reportant la succession de lettre présentes sur son génome, il nous faudrait plus de 2 millions de pages comme celle ci.

C'est grâce à ce code génétique que les cellules se divisent, se multiplient et se spécialisent pour donner naissance à ces organes aussi divers que le cœur, les poumons ou le cerveau. Les maladies génétiques sont directement liées à la manière dont ces textes sont écrits.

Le code génétique de l'homme, maintenant entièrement transcrit, est même disponible en accès libre sur Internet. Celui d'autres espèces est en cours de décodage. La compréhension de ce message de plusieurs millions de codes donne de nouvelles perspectives à la recherche médicale. Le décodage du génome des plantes ouvre de nouvelles perspectives dans le domaine de l'agriculture.

L'exploitation de bases de données génomiques ne peut s'envisager sans le *data mining*.

16. Les logiciels de *data mining*

Au début des années 90, les logiciels estampillés *data mining* se faisaient rares Ils étant uniquement l'apanage de petites entreprises novatrices qui implémentaient des méthodes directement issues des thèses de doctorat développées dans les laboratoires de recherche. Ces outils, somme toute assez sommaires, quand ils n'étaient pas que des assemblages de bibliothèques de programmes, se spécialisaient sur une méthode ou une variété de méthodes appartenant au même paradigme, peu connus en statistiques, mais très en vogue au sein de la communauté de l'apprentissage automatique et de la reconnaissance des formes. Leur diffusion demeurait assez restreinte, même dans le milieu de la recherche.

Le vrai décollage est survenu dans le milieu des années 90. A cette période, les petits logiciels ont pu accéder à des interfaces professionnelles avec le développement des solutions *data mining* sous des environnements Windows. Les petites sociétés ont alors commencé à toucher de manière significative le marché des entreprises, plusieurs études montrèrent des perspectives très optimistes, laissant à penser que le besoin futur en outils de *data mining* était un filon à très haut rendement.

L'offre d'outils de *data mining* est aujourd'hui pléthorique. Rien qu'en faisant une recherche sur Internet avec les mots clés « software » et « *data mining* »,

il faudrait plusieurs jours pour dépouiller manuellement les résultats, signe de l'importance prise par ce domaine.

17. Bibliographie

Dans cette bibliographie, nous avons essentiellement inséré seulement les ouvrages de base. Les articles de revues ou des conférences ont été explicitement écartés. On peut trouver sur Internet des bibliographies assez larges sur les différents sujets.

J-P. Auray, G. Duru, A. Zighed; Analyse des données multidimensionnelles, volume 2 : les méthodes de structuration ; A. Lacassagne ; 2000.

J-P. Auray, G. Duru, A. Zighed ; Analyse des données multidimensionnelles, volume 3 : les méthodes d'explication ; A. Lacassagne ; 2000.

J-P. Auray, G. Duru, A. Zighed ; Analyse des données multidimensionnelles, volume 1 : les méthodes de description ; A. Lacassagne ; 2000

A. Agresti ; An Intruduction to Catagorical Data Analysis ; John Wiley & Sons ; 1996

P. Baldi and . Brunak ; Bioinformatics: The Machine Learning Approach ; MIT Press ; 1998

S. Benninga et B. Czaczkes ; Financial Modeling ; MIT Press ; 1997

J. Bertin ; Graphics and Graphic Information Processing ; Berlin ; 1989

L. Breiman, J. Friedman, R. Olshen et C. Stone ; Classification and Regression trees ; Wadsworth International Group ; 1984

M. Berthold et D. J. Hand ; Intelligent Data Analysis: An Introduction ; Springer-Verlag ; 1999

C. M. Bishop ; Neural Network for Pattern Recognition ; Oxford University Press ; 1995

V. Barnett et T. Lewis ; Outliers in Statistical Data ; John Wiley & Sons ; 1994

M. J. A. Berry et G. Linoff ; Mastering Data Mining: The Art and Science of Customer Relationship Management ; John Wiley & Sons ; 1999

A. Baxevanis et B. F. F. Ouellette ; Bioinformatics: A pratical Guide to the Analysis of Genes and Proteins ; John Wiley & Sons ; 1998

A. Berson et S.J. Smith ; Data Warehousing, Data Mining, and OLAP ; McGraw-Hill ; 1997

A. Berson, S. J. Smith et Thearling ; Building Data Mining Applications for CRM ; McGraw-Hill ; 1999

R. Baeza-Yates et B. Ribeiro-Neto ; Modern Information Retrieval ; Addison-Wesley ; 1999

Z. Chen ; Intelligent Data Warehousing: From Data Preparation to Data Mining ; CRC Press ; 2001

K. Cios, W. Pedrycz et R. Swiniarski ; Data Mining Methods for Knowledge Discovery ; Kluwer Academic Publishers ; 1998

Y. Chauvin et D. Rumelhart ; Backpropagation: Theory, Architectures and Applications ; Hillsdale, NJ:LawrenceErlbaum Assc. ; 1995

R. Duda and P. Hart ; Pattern Classification and Scene Analysis ; John Wiley & Sons ; 1973

D. J. Hand, H. Mannila et P. Smyth ; Principles of Data Mining ; MIT Press ; 2001

M. Dash et Liu ; Feature Selection Methods for Classification ; Intelligent Data Analysis: An International Journal ; 1997

A. J. Dobson ; An Introduction to Generalized Lineair Models ; Chapman and Hall ; 1990

R. Elmasri et S.B. Navathe ; Fundamentals of Data bases systems ; Benjamin/Cummings ; 1994

B. Efron et R. Tibshirani ; An Introduction to the Bootstrap ; Chapman and Hall ; 1993

Usama Fayyad, Georges Grinstein, Andreas Wierse ; Information Visualization in Data Mining and Knowledge Discovery ; Morgan Kaufmann ; 2001

U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth et R. Uthurusamy (Eds.) ; Advances in Knowledge Discovery and Data Mining ; MIT Press ; 1996

L. Fu ; Neural Networks in Computer Intelligence ; McGraw-Hill ; 1994

R. Groth ; Data Mining: Building Competitive Advantage ; Prentice Hall ; 1999

J. A. Hartigan ; Clustering Algorithms ; John Wiley & Sons ; 1975

C.H. Huberty ; Applied Discriminant Analysis ; John Wiley & Sons ; 1994

W. H. Inmon ; Building the Data Warehouse ; John Wiley & Sons ; 1996

M. James ; Classification Algorithms ; John Wiley & Sons ; 1984

M. Jambu ; Méthodes de l'analyse des données ; Eyrolles ; 1999

M. Jambu ; Introduction au Data Mining ; Eyrolles ; 1999

A. K. Jain et R. C. Dubes ; Algorithms for Clustering Data ; Prentice Hall ; 1988

F.V. Jensen ; Introduction to Bayesien Networks ; Springer-Verlag ; 1996

R. A. Johnson et D.A. Wichern ; Applied Multivariate Statistical Analysis ; Prentice Hall ; 1992

R. L. Kennedy, Y. Lee, B. Van Roy, C.D. Reed et R. P. Lippman ; Solving Data Mining Problems Through Pattern Recognition ; Prentice Hall ; 1998

J.L. Kolodner ; Case-Based reasoning ; Morgan Kaufmann ; 1993

R. Kimball ; The Data Warehouse Toolkit ; John Wiley & Sons ; 1996

P. Langley ; Elements of Machine Learning ; Morgan Kaufmann ; 1996

H. Liu et H. Motoda (Eds) ; Feature Extraction, Construction and Selection: A data Mining Perspective ; Kluwer Academic Publishers ; 1998

H. Liu et H. Motoda ; Feature Selection for Knowlede Discovery and Data Mining ; Kluwer Academic Publishers ; 1998

René Lefébure, Gilles Venturi ; Data Mining ; Eyrolles ; 2001

R. Mattison ; Data Warehousing and Data Mining for Telecommunications ; Artech House ; 1997

R. S. Michalski, I. Bratko et M. Kubat ; Machine Learning and Data Mining: Methods and Applications ; John Wiley & Sons ; 1998

J. Mena ; Data Mining Your Website ; 2001

T. M. Mitchell ; Machine Learning ; McGraw-Hill ; 1996

P. Naim et M. Bazsalicza ; Data mining pour le Web ; Eyrolles ; 2001

J. Pearl ; Probabilistic Reasoning in Intelligent Systems ; Morgan Kaufmann ; 1988

G. Piatetsky-Shapiro et W. J. Frawley ; Knowledge Discovery in Databases ; MIT Press ; 1991

D. Pyle ; Data Preparation for Data Mining ; Morgan Kaufmann ; 1999

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

D.E. Rumelhart, J.L. McClelland ; Parallel Distributed Processing ; MIT Press ; 1986

J.W. Shavlik et G.T. Dietterich ; Readings in Machine Learning ; Morgan Kaufmann ; 1990

C. Seidman ; Data Mining with Microsoft SQL Server 2000 Technical Reference ; IT Prodessinal Microsoft ; 2001

D. Sullivan ; Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing, and Sales ; John Wiley & Sons ; 2001

E. Thomsen ; OLAP Solutions: Building Multidimensional Information Systems ; John Wiley & Sons ; 1997

B. Thuraisingham ; Data Mining : Technologies, Techniques, Tools, and Trends ; 2001

J. M. Tuffery ; Data Mining et scoring, Bases de données et gestion de la relation client ; Dunod ; 2002

E.R. Tufte ; The Display of Quantitative Information ; Graphics Press ; 1983

E.R. Tufte ; Visual Explanations: Images and Quantities, Evidence and Narrative ; Graphics Press ; 1997

I. H. Witten et E. Frank ; Data Mining ; Morgan Kaufmann ; 1999

S. M. Weiss et N. Indurkhya ; Predictive data mining ; Morgan Kaufmann ; 1998

D.A. Zighed R. Rakotomalala ; Graphes d'induction et Data Mining ; Hermès ; 2000

A. Zighed, G. Duru, J-P. Auray ; Sipina, méthode et logiciel ; A. Lacassagne 2000