MENTION TRÈS HONORABLE
avec félicitations du jury

UNIVERSITE DE NANTES

FACULTE DE MEDECINE

Année 2002-2003 N°

THESE

Pour le

DIPLOME D'ETAT DE DOCTEUR EN MEDECINE

QUALIFICATION EN MEDECINE GENERALE

Par

Antoine COUPVENT des GRAVIERS

Présentée et soutenue publiquement le 24 juin 2003

UTILISATION D'UNE TECHNIQUE D'APPRENTISSAGE NON PARAMETRIQUE SUR UNE BASE DE DONNEES DE 87 ENFANTS ATTEINTS DE NEUROFIBROMATOSE DE TYPE 1

Président : M. Le Professeur J-F. STALDER

SOMMAIRE

INTRODUCTION

La masse des connaissances médicales en général et le volume des bases de données médicales en particulier ne cessent d'augmenter à un rythme que certains qualifient d'exponentiel.

Et lorsque, après avoir collecté consciencieusement des données médicales durant des mois et des années, le médecin se retourne et contemple son œuvre, il se peut qu'effrayé par la montagne de giga octet emmagasiné, il se tourne vers le statisticien. Il n'est pas rare (et c'est un euphémisme) que celui-ci lui réponde au meilleur des cas qu'il ne pourra répondre que partiellement à la question que le médecin entendait résoudre avec ses données, au pire qu'il est impossible d'exploiter sa base...

Et pourtant le médecin, en expert de la matière qui l'avait conduit à créer et remplir sa base de données, sait intuitivement qu'il y existe de l'information et qu'il suffirait de savoir l'extraire correctement.

Face à ce problème qui, rassurons nous, n'existe pas seulement dans la matière médicale, des algorithmiciens ont cherché à développer des méthodes de datamining (première partie), d'extraction de connaissances, toujours différentes et plus performantes.

Parmi elles l'apprentissage non paramétrique, c'est-à-dire l'apprentissage qui ne part pas d'un modèle construit à priori, nous a semblé pouvoir s'appliquer particulièrement bien aux données médicales.

A partir d'une base de données de 87 enfants atteints de neurofibromatose de type 1, utilisée par le Docteur Sébastien Barbarot[1],et des premiers enseignements qu'il en avait tirés (deuxième partie), nous avons cherché à mettre à l'épreuve une technologie d'intelligence artificielle par apprentissage non paramétrique sur ces mêmes données.

Plusieurs hypothèses relatives à la neurofibromatose de type 1 ont été définies à cet effet : Les enfants ayant des complications de types difficultés d'apprentissage ont-ils un profil particulier par rapport aux enfants ne présentant pas de telles difficultés d'apprentissage ? Existe-t-il des règles de corrélations entre la présence de tel ou tel symptôme et l'existence d'une pseudarthrose ? Existe-t-il des règles de corrélations entre différents symptômes de la maladie et le sexe de l'enfant atteint ?(troisième partie).

[1] BARBAROT Sébastien, NEUROFIBROMATOSE DE TYPE 1 : INCIDENCE ET PREVALENCE DES COMPLICATIONS CHEZ 87 ENFANTS ATTEINTS, Thèse de Doctorat en Médecine, soutenue le 26 octobre 1999, Faculté de Médecine de Nantes

PREMIERE PARTIE :

INTRODUCTION AU DATA MINING ET AUX TECHNOLOGIES D'INTELLIGENCE ARTIFICIELLE PAR APPRENTISSAGE NON PARAMETRIQUE.

I. DEFINITION GENERALE DU DATA MINING

Le Data Mining est un processus d'aide à la décision où les utilisateurs cherchent des modèles d'interprétation dans les données.

Le Data Mining a été (et est toujours) principalement utilisé en marketing (segmentation de clientèle, analyse de paniers d'achats, profiling, ...) et en finance (détection de fraude, évaluation du risque, prédiction de cours, ...), mais également dans l'industrie de production.
Ces domaines se caractérisent entre autres par la présence de très grandes quantités de données, la possibilité de faire intervenir des experts (statisticiens, experts du data mining) dans le processus d'analyse des données, la grande fiabilité (due à l'avancement des systèmes d'information) et complétude des bases de données

L'émergence des « Data Warehouse », qui concentrent des données en provenance d'applications CRM de l'entreprise, a encore contribué à attirer l'attention de la communauté du Data Mining vers les classes de problèmes ayant de très fortes volumétries (typiquement le traitement de tera-octets de données). Les applications plus récentes de « Webmining » (analyse de logs sur des applications Internet) ont également attiré l'attention ces dernières années vers l'analyse de quantités de données toujours plus grandes.

Un fort besoin existe dans l'industrie d'outils permettant d'analyser les données de conception et de production bien moins volumineuses, afin d'améliorer les produits

(leur conception initiale et leur qualité), et de maximiser la productivité des outils de production. Ce besoin n'est pas aujourd'hui pleinement satisfait

Le profit attendu de ces outils est l'amélioration de la structure de coûts de l'entreprise et une meilleure réactivité.

Les méthodes de résolution de problèmes (DMAIC par exemple), les techniques de MSP (Maîtrise Statistique des Procédés), et leur systématisation dans des approches telles que « 6 Sigma », visent à obtenir une qualité de production constante en appliquant une méthodologie systématique de résolution de problème. Ces approches permettent de garantir que les problèmes rencontrés par l'entreprise sont abordés avec méthode, et que leur solution n'est pas ponctuelle mais au contraire durable.

Si on se réfère à la méthode « 6 Sigma », on décompose la résolution de problème en les étapes suivantes : Définir, Mesurer, **Analyser**, **Améliorer**, Contrôler, **Standardiser**. L'apport attendu du Data Mining dans ces approches se situe :

- Dans les étapes « Analyser », « Améliorer », où il peut permettre d'identifier rapidement des solutions potentielles à un problème déjà bien posé et dont on sait que l'on saura pérenniser cette solution.
- Dans l'étape « Standardiser », où des déclinaisons « temps réel » de techniques de data mining peuvent être utilisées pour effectuer automatiquement le réglage de certaines machines.

Le déploiement par exemple de la MSP dans les ateliers ou la mise en œuvre des plans d'expérience impliquent des contraintes lourdes de coût ou d'organisation. Il existe un besoin latent d'outils permettant d'alléger ces procédures dans un cadre de production.

Le Data Mining devrait apporter une réponse à ces besoins, mais n'est pas parvenu dans le passé à satisfaire massivement les besoins des industriels. Les problèmes

industriels se prêtent en effet aujourd'hui difficilement à l'utilisation de l'approche classique du data mining pour un ensemble de raisons détaillées ci-dessous.

Nature des compétences disponibles – les compétences nécessaires à la mise en œuvre opérationnelle de la plupart des techniques de data mining ne sont pas présentes dans les environnements de production. Les personnels de production sont en effet en général compétents en statistiques de base (écarts types, moyennes) nécessaires à l'application des techniques de MSP par exemple ou en statistiques avancées dans le cas où une démarche « 6 sigma » aurait été déployée dans l'entreprise, mais ont très rarement des connaissances en data mining.

Données peu nombreuses – les données disponibles sur les défauts ou les pannes en production sont parfois peu nombreuses (les produits défectueux sont moins nombreux que les produits de bonne qualité, les pannes se produisent rarement), et sont souvent coûteuses à acquérir (contexte de tests ou essais, hors production). On préférera donc souvent des techniques permettant une action rapide en présence d'une petite quantité de données à des techniques exigeant pour être appliquées la collecte d'une grande quantité de données.

Données fragmentaires – malgré l'idéal du CIM (Computer Integrated Manufacturing) qui présente la production comme devant être totalement intégrée au système d'information de l'entreprise, et aux normes ISO 9000 de contrôle qualité, les usines sont aujourd'hui constituées de ce qu'il convient d'appeler des « îlots d'automation », dans lesquels l'information disponible est incomplète. De nombreuses données qui pourraient être utiles ne sont pas mesurées, sont répertoriées manuellement ou le sont depuis peu de temps.

Priorité à l'explication – la capacité explicative d'une technique est plus précieuse pour tirer les enseignements des explications extraites et les pérenniser au niveau de la production que la capacité prédictive des algorithmes utilisés. De plus, elle s'insèrera parfaitement dans les démarches DMAIC ou « 6 Sigma ».

Données peu fiables – la présence de mesures ou relevés manuels (et donc peu fiables) n'est pas un phénomène destiné à disparaître dans les années qui viennent. En effet, le coût associé à l'automatisation des mesures est souvent élevé, et certaines mesures sont parfois impossibles pour les nombreux relevés qualitatifs. Des techniques capables d'utiliser ces données telles qu'elles sont aujourd'hui, et d'en extraire le maximum d'information apportent plus de valeur à l'industrie.

Priorité au résultat – la preuve de la validité d'une méthode est démontrée pour un industriel par le résultat obtenu. Contrairement au marketing, où il est difficile d'établir si les décisions prises ont les conséquences attendues, il est facile en production de constater l'effet de nouveaux réglages sur le taux de rebuts d'un atelier de production. Cette capacité à produire un retour d'expérience rapide encourage donc fortement les techniques aux résultats tangibles. A contrario, le fait qu'une méthode s'appuie sur des principes théoriques bien établis n'est pas en soi un avantage décisif.

Fortes contraintes de délai – on préfère souvent une solution approchée mais rapide aux problèmes de production qu'une solution optimale qui prendrait beaucoup de temps à être mise en œuvre. La flexibilité aujourd'hui nécessaire à la production ainsi que les coûts engendrés par une réaction trop lente aux problèmes favorisent les approches nécessitant moins de temps.

Paramètres hétérogènes et nombreux – la compréhension des procédés industriels implique souvent un nombre de paramètres important (plusieurs

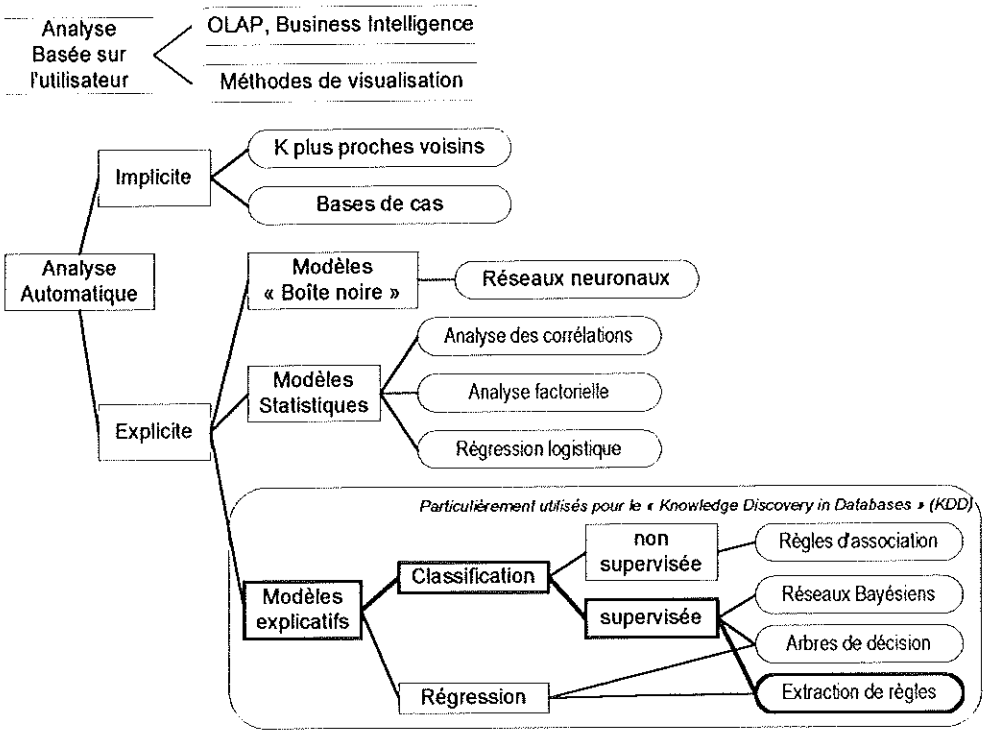
centaines), et de natures différentes : quantitatives et qualitatives. Les techniques ne fonctionnant que pour un petit nombre de paramètres ou prenant en compte des variables uniquement quantitatives ou uniquement qualitatives présentent un intérêt très limité pour les industriels.

Dans le domaine médical, certes, un certain nombre de contraintes spécifiques à l'industrie ne vont pas être retrouvé. Cependant que ce soit en terme de volume de données (énorme base de données difficile à « manier » ou au contraire échantillon de patients très faible), d'imprécision des données, d'hétérogénéité des données, et surtout d'évolution de la connaissance médicale... les bases de données médicales vont ressembler aux bases de données industrielles et l'ensemble des idées nouvelles et des mécanismes de Data Mining expérimentés sur les bases de données industrielles peuvent également l'être sur les bases médicales.

II. LE DATA MINING : UN LARGE PANEL DE TECHNIQUES

Les techniques utilisées en analyse de données sont nombreuses, et les outils les plus populaires sont des « Boîtes à outils » dans lesquelles la responsabilité du choix des techniques adaptées au problème à traiter est laissée à l'utilisateur. Il existe de nombreuses approches pour les caractériser.

Le graphique suivant propose une classification de techniques existantes parmi les plus utilisées.



Basées sur l'utilisateur / automatiques – De nombreuses techniques d'analyse de données, en particulier les techniques dites OLAP (On Line Analytical Processing) ou de « Business Intelligence », ainsi que d'autres techniques se basant sur la visualisation multidimensionnelle des données, font appel à l'utilisateur pour extraire lui-même l'information pertinente des données. Elles se limitent donc à lui faciliter

cette recherche. Au contraire, les techniques de « Data Mining » à proprement parler cherchent à extraire automatiquement un modèle à partir des données disponibles.

Implicites / explicites – Parmi les techniques automatiques, les techniques implicites (par exemple les « k plus proches voisins » ou les techniques de raisonnement à base de cas) utilisent les données elles-mêmes pour effectuer des prédictions, alors que les techniques explicites utilisent un modèle des données, quelle que soit sa forme (par exemple, un arbre de décision, des règles, un réseau neuronal).

Modèles « boîte noire » / explicatifs - Parmi les techniques explicites, certaines utilisent un modèle de représentation des données ne permettant pas d'explication (par exemple un réseau de neurones). Ces modèles sont dits des modèles « boîte noire ». Les modèles pouvant fournir une explication des données analysées sont qualifiés de « Boîte blanche », de modèles explicatifs ou de modèles de connaissances. Ils permettrent de contribuer à la compréhension du système étudié et sont donc utilisés pour mettre en œuvre le processus d'extraction automatique de connaissances (ou « KDD » pour Knowledge Discovery in Databases). C'est le cas notamment des modèles basés sur des règles, des arbres de décision ou des réseaux Bayésiens.

Classification supervisée / non supervisée / régression – Les méthodes de classification supervisée permettent de prédire ou d'expliquer une variable prenant un certain nombre de valeurs discrètes connues à l'avance (les classes) à partir d'autres variables discrètes ou continues. Les méthodes de classification non supervisée ou de « clustering » permettent de regrouper les données disponibles en un certain nombre de classes non connues à l'avance. Les techniques de régression permettent de prédire ou d'expliquer une variable continue à partir d'autres variables

continues ou discrètes. Le plus souvent, les problèmes à résoudre dans un environnement industriel de production sont bien définis. Par exemple, on cherche à minimiser le taux de rebuts, à minimiser le niveau de pollution d'un processus, à maximiser le volume de fabrication d'une ligne. Dans ce contexte, les techniques de classification supervisée et de régression sont utiles, mais les techniques de classification non supervisée le sont beaucoup moins.

Quels sont les avantages et les inconvénients de ces différentes méthodes ?

Analyse par l'utilisateur : Les techniques OLAP sont essentiellement destinées à exploiter efficacement de très grandes bases de données en temps réel. Les techniques de visualisation des données, elles, sont utilisées sur des échantillons de données moins importants. Elles sont souvent utiles aux utilisateurs industriels pour se faire une idée sur les données à traiter et pour repérer les échantillons de données contenant des erreurs. Elles s'utilisent souvent en préalable à une autre technique.

Techniques implicites : Ces techniques sont largement utilisées en diagnostic (bases de cas) ou en reconnaissance des formes (k plus proches voisins), dans des cas d'utilisation où on ne requiert pas d'explication mais seulement une prédiction. Elles ne sont pas adaptées à des bases présentant beaucoup de données manquantes, car elles s'appuient sur des calculs de distances.

Réseaux neuronaux : Les réseaux neuronaux sont largement utilisés pour des problèmes pour lesquels on ne requiert pas d'explication mais seulement une prédiction (par exemple la reconnaissance des formes, le diagnostic). Ils ne sont pas adaptés à des bases présentant beaucoup de données manquantes, car les

techniques d'apprentissage utilisées requièrent la connaissance complète des échantillons. Leur utilisation demande des compétences spécifiques (réglage du nombre de neurones, nombres de couches, dynamique de l'apprentissage), peu compatibles avec une utilisation massive.

Analyse factorielle, analyse des corrélations : L'analyse factorielle permet de découvrir certaines combinaisons de variables ayant une incidence forte sur la variable que l'on cherche à prédire. Ces combinaisons sont cependant en général difficiles à interpréter, et leur usage nécessite une compétence certaine. L'analyse des corrélations, elle, est plus largement utilisée dans l'industrie car elle permet de comprendre quelles variables sont très corrélées entre elles, et est donc utilisée pour simplifier le problème à analyser avant d'utiliser une autre technique.

Réseaux Bayésiens : Les réseaux Bayésiens permettent de modéliser des chaînes d'évènements avec des relations probabilistes de cause à effets. Ils permettent de prendre en compte une expertise à priori, et il est possible d'interpréter les résultats obtenus par l'apprentissage. Cependant, cette interprétation ainsi que la mise en œuvre exige une forte compétence en statistiques.

Règles et Arbres de décision : Les arbres de décision et les règles permettent l'interprétation des résultats la plus naturelle parmi toutes les technologies de Data Mining. L'apprentissage des règles, à priori plus intéressant de ce point de vue (voir le paragraphe suivant), était jusqu'à maintenant un problème difficile, ce qui explique le manque d'offres de ce type. Les approches d'apprentissage d'arbres de décision sont par contre bien documentées et assez largement utilisées malgré leurs limitations.

III. LE CAS SPECIFIQUE DE L'EXTRACTION AUTOMATIQUE DE REGLES

Les règles et les arbres de décision sont deux approches de représentation de la connaissance qui semblent au premier abord similaires, mais qui diffèrent cependant fortement. On peut considérer que les arbres de décision sont une forme simplifiée des règles.

Le principe des arbres de décision est de couper la base par des tests logiques successifs, organisés sous la forme d'un arbre. Il existe diverses approches pour aboutir à ce découpage mais on aboutit toujours à un arbre comme par exemple celui donné ci-dessous. Ce principe implique que l'on accorde une importance particulière aux variables utilisées comme premiers nœuds de l'arbre.

Tout arbre de décision peut être exprimé sous forme de règles. La lecture d'un arbre de décision en vue d'une explication passe en général par cette transcription sous forme de règles. Ces règles ne peuvent pas se chevaucher et constituent une partition de l'espace. Le nombre de prédicats par règle augmente rapidement avec la profondeur de l'arbre, et les mêmes prédicats apparaissent dans plusieurs, voire toutes les règles (pour la première condition du haut de l'arbre). Le système de règles généré est en général moins lisible que le système de règles optimal équivalent.

Exemple : évaluation du confort thermique

Cet exemple didactique montre comment des règles et un arbre de décision peuvent être utilisés pour évaluer le confort thermique. On cherche avec un modèle le plus simple possible à évaluer en fonction de la température et l'humidité de l'air les

zones de confort et d'inconfort. La zone de confort est en vert (foncé) et la zone d'inconfort en beige (clair).

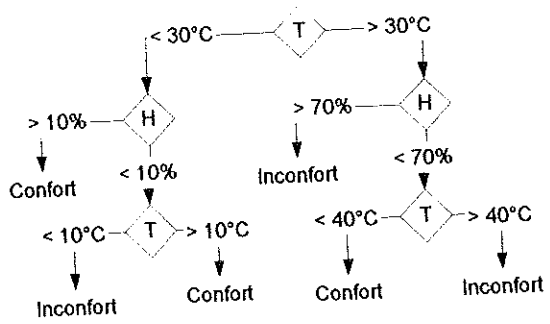
Décision sans zones d'abstention

On cherche à mettre au point un classifieur capable de décider pour toute valeur de la température et de l'humidité si on se trouve dans une zone de où l'on se sent confortable (classe « Confort ») ou inconfortable (classe « inconfort »).

Arbre de décision

Utilisons un arbre de décision pour résoudre ce problème. Il est possible de construire l'arbre ci-dessous comportant 5 nœuds, qui peut alors être traduit pour l'interpréter sous la forme de 6 règles totalisant 16 prédicats.

Sur le schéma montrant les zones de décision en fonction de la température et de l'humidité, les chiffres (1), (2) et (3) correspondent aux niveaux de l'arbre de décision.



R1 SI T < 30°C
ET H > 10%
ALORS Confort

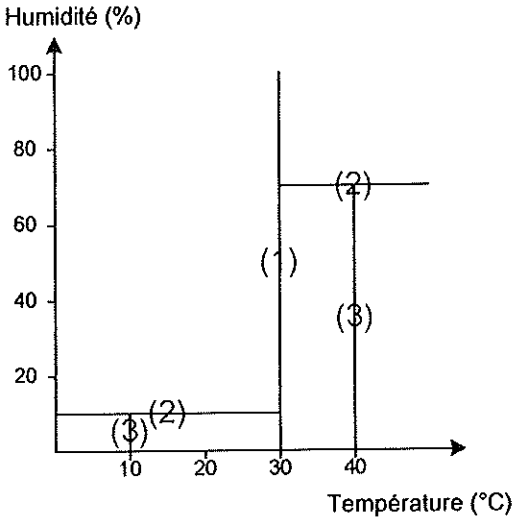
R2 SI T < 30°C
ET H < 10%
ET T < 10°C
ALORS Inconfort

R3 SI T < 30°C
ET H < 10%
ET T > 10°C
ALORS Confort

R4 SI T > 30°C
ET H < 70%
ALORS Inconfort

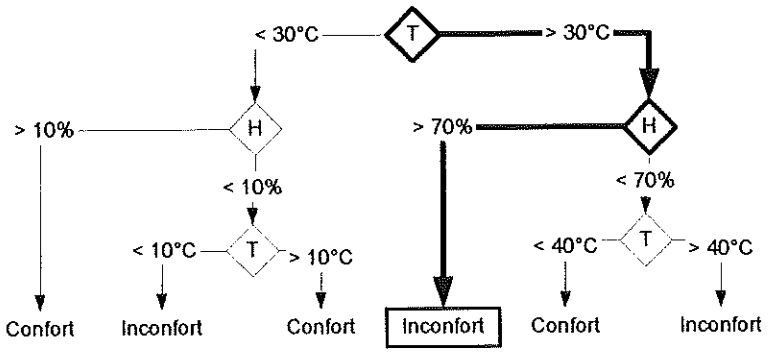
R5 SI T > 30°C
ET H < 70%
ET T < 40°C
ALORS Confort

R6 SI T > 30°C
ET H < 70%



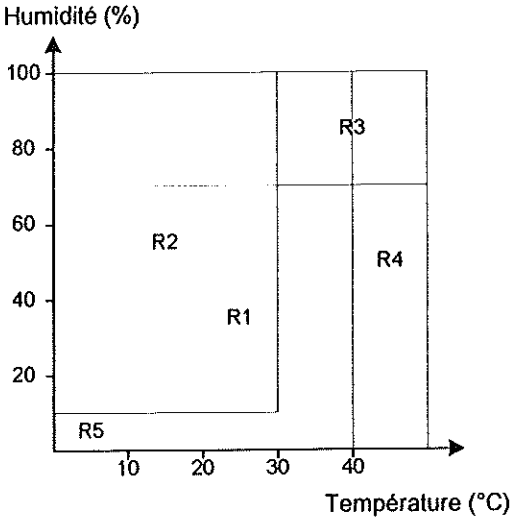

```
ET      T > 40°C
ALORS   Inconfort
```

Pour effectuer une prédiction, on parcourt l'arbre de haut en bas, en effectuant l'un après l'autre les tests correspondant à chaque nœud parcouru. Par exemple, si on a Température = 33°C et Humidité = 80%, on parcourt l'arbre comme le montre la figure ci-dessous :



Règles avec mécanisme de vote

Avec un modèle basé sur des règles, il est possible d'expliquer (voir ci-dessous) le phénomène avec 5 règles qui totalisent 10 prédicats. On remarque que les règles peuvent se chevaucher.




```
R1   SI      T > 10°C
     ET      T < 40°C
     ET      H < 70%
     ALORS   Confort

R2   SI      T < 30°C
     ET      H > 10%
     ALORS   Confort

R3   SI      T > 30°C
     ET      H > 70%
     ALORS   Inconfort

R4   SI      T > 40°C
     ALORS   Inconfort

R5   SI      T < 10°C
     ET      H < 10%
     ALORS   Inconfort
```

Pour effectuer une prédiction, on détermine à partir des entrées quelles règles sont vraies. On calcule alors la sortie à l'aide d'un mécanisme de vote entre les règles. Ce mécanisme intervient quand plusieurs règles vraies en même temps ont pour conclusion des classes différentes. Ici, les règles de la classe « Confort » et celles de la classe « Inconfort » ne se chevauchant pas, et il n'est donc pas nécessaire de faire intervenir ce mécanisme.

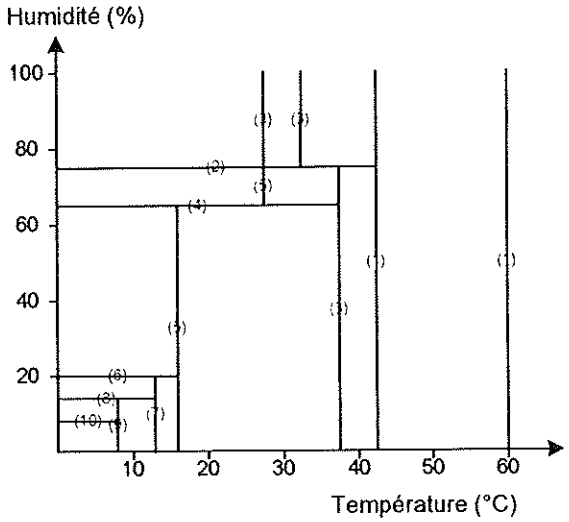
Par exemple, si « Température = 33°C » et « Humidité = 80% », seule la règle R3 est vraie, et on se trouve donc avec comme décision « Inconfort ».

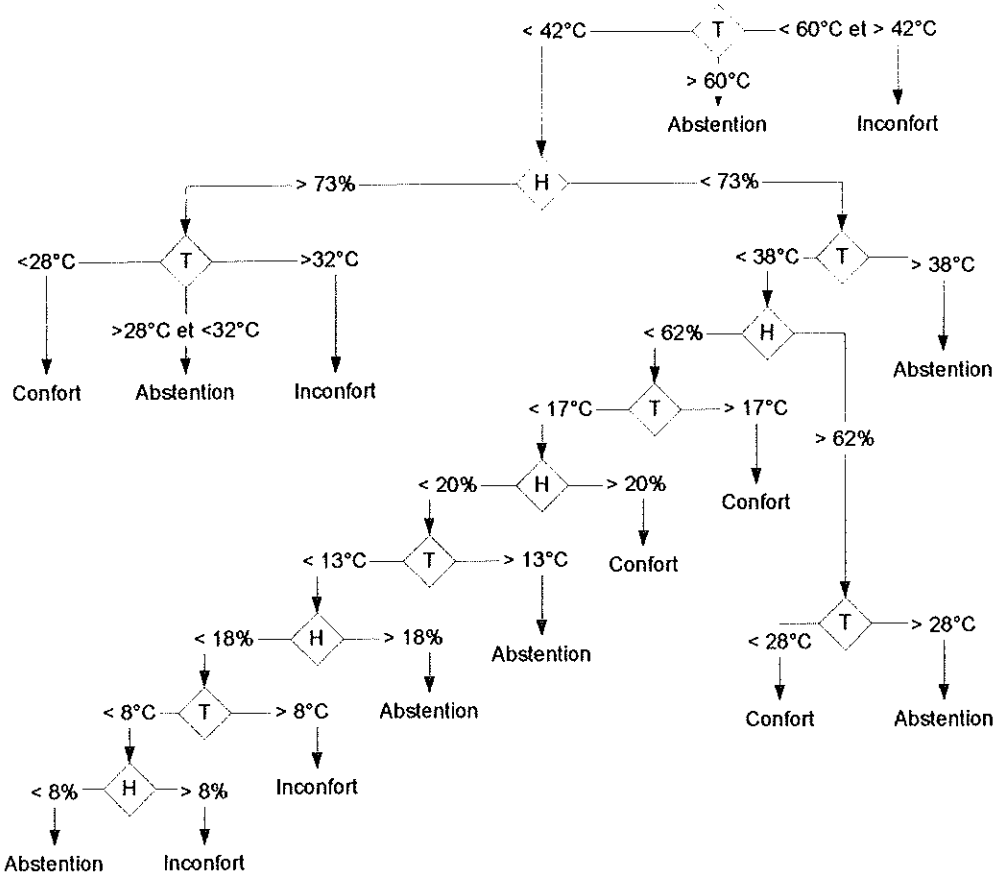
Décision avec zones d'abstention

On considère maintenant qu'il existe une zone intermédiaire (en gris moyen sur les schémas suivants) où on n'est ni confortable ni inconfortable. D'autre part, on souhaite que les zones plus qu'inconfortables (pour lesquelles la personne humaine est en danger) ne fassent pas partie de la classe « Inconfort ».

Arbre de décision

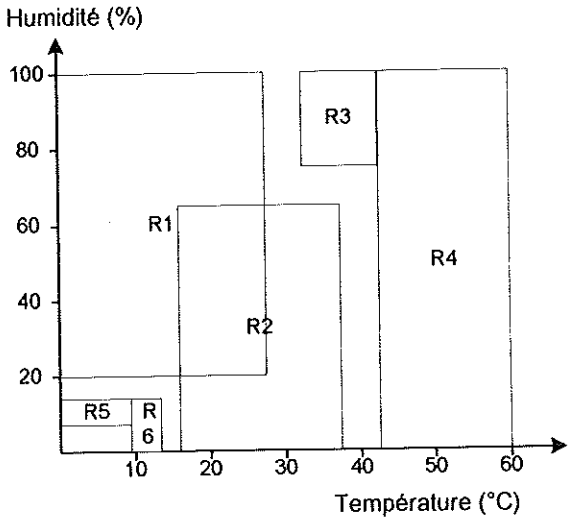
Avec un arbre de décision, il est nécessaire de créer une classe supplémentaire « Abstention » qui modélise le fait que l'on s'abstient de prendre une décision dans certains cas.
On constate qu'il est possible de prendre en compte cette nouvelle classe avec un arbre de décision comportant 14 nœuds.
Cet arbre de décision peut être traduit sous forme de 15 règles comportant au total 67 prédicats.





Règles avec mécanisme de vote

Dans le cas d'un système de règles, il suffit de 6 règles totalisant 12 prédicats. Le système de règles produit une abstention dans les zones pour lesquelles aucune règle n'est active.
Dans les zones où des règles correspondant à des classes différentes se chevauchent (ce qui n'est pas le cas sur cet exemple), un mécanisme de vote majoritaire entre les règles (avec une marge de décision) permet d'effectuer la décision, et éventuellement de s'abstenir.



Conclusions

On comprend sur cet exemple que les modèles basés sur l'extraction automatique de règles sont plus simples et les règles plus lisibles que ceux fournis par les arbres de décision. Ce phénomène est plus marqué pour des problèmes complexes.

De manière générale, des règles construites à l'aide de cette technologie peuvent se chevaucher ou au contraire ne couvrir qu'une partie de l'espace (ce qui est normal, car dans beaucoup de problèmes certaines portions de l'espace sont physiquement impossibles).

Par contre, les règles obtenues à partir d'un arbre de décision sont exclusives et couvrent tout l'espace (on dit qu'elles définissent une partition de l'espace). Cette propriété contribue à expliquer :

- Le fait que les règles produites à partir des arbres de décision ne puissent pas identifier toutes les règles qui seraient intéressantes : certaines en excluent d'autres car leur somme viole la contrainte de partitionnement de l'espace
- Le fort pouvoir descriptif et explicatif constaté en général avec des règles par rapport aux arbres de décision

- Les meilleurs résultats en prédiction obtenus pour des systèmes de règles avec vote, avec un grand nombre de règles se chevauchant (par construction, l'effet statistique du au vote ne peut pas être reproduit avec un arbre de décision). Ainsi, en jouant sur le nombre de règles se recouvrant, il est possible de s'adapter à l'objectif de l'utilisateur (explication ou prédiction).

D'autre part, un jeu de règles votant entre elles définit implicitement une distance non paramétrique à comportement local. Cette distance est ajustée par l'algorithme d'apprentissage de Pertinence Data Intelligence en fonction des données de la base et de leur qualité.

Egalement, les bases de règles avec système de vote entre les règles permettent de gérer la possibilité pour le système de s'abstenir dans les cas où la conclusion n'est pas certaine. Cet aspect est important si une mauvaise décision peut avoir de fortes incidences de coût ou impacter la sécurité.

IV. COMPARAISON DU DATA MINING ET DES STATISTIQUES

Les techniques statistiques utilisées en analyse de données sont principalement basées sur des tests statistiques d'hypothèses. Leur force dérive de leur validité scientifique. En général, la tradition scientifique en statistiques veut qu'une méthode soit publiée et utilisée si la preuve peut être faite de sa validité. Cette attitude s'oppose à la tradition scientifique de l'informatique et du Data Mining, pour lesquels prévaut une démarche expérimentale. Il en résulte que de nombreux algorithmes de Data Mining donnent de bons résultats en pratique sans que des démonstrations formelles de leur validité existent.

Un autre point commun des approches statistiques est l'importance donnée à la capacité de généraliser des conclusions à toutes les données à partir d'un petit échantillon de ces données. Cependant, les problèmes auxquels s'attaque le Data Mining se caractérisent souvent par le fait que toutes les données disponibles doivent être utilisées par l'apprentissage, sous peine de perdre de l'information, en particulier quand les échantillons de données sont relativement peu nombreux (par exemple des défauts se produisant rarement, des fraudes, ...).

Enfin, les méthodes statistiques partent pour la plupart du principe que l'on cherche à valider un modèle donné décidé à priori. Ces techniques sont qualifiées d' « apprentissage paramétrique », où l'on dispose d'un bon modèle des données, et où l'apprentissage consiste à caler le modèle, donc à estimer les valeurs optimales des paramètres.

Au contraire, le processus de Data Mining est un apprentissage « non paramétrique » par essence exploratoire : il se préoccupe de la possibilité de découvrir des connaissances inattendues et de valeur dans les données, alors que

les tests d'hypothèse, par nature, visent à confirmer ou infirmer une connaissance à priori. Ainsi, on peut penser que les statistiques permettront à terme de confirmer ou d'infirmer les connaissances mises en lumière par les procédures de Data Mining. Finalement, un processus exploratoire consiste à découvrir des connaissances inconnues, donc imprévues à priori. Un tel processus, par nature, se déroule généralement en plusieurs étapes, en fonction d'objectifs redéfinis à chaque étape en fonction de ce qui a été découvert. Ce processus correspond bien à méthodologie du data mining.

V. FONCTIONNEMENT DU LOGICIEL DE LA SOCIETE PDI

La société PDI développe une technologie d'extraction automatique de règles. Ces règles permettent de faire de la classification supervisée.

La technologie développée par PDI permet d'effectuer l'apprentissage de règles avec un mécanisme de vote. Ce type de base de règles présente les avantages suivants :

- Forte capacité à expliciter (les règles simples à lire et comprendre)
- Très bons résultats prédictifs (grâce à la possibilité de recouvrement des règles)

Les méthodes ascendantes utilisent des comportements locaux et les agrègent afin de dégager des comportements globaux, alors que les méthodes descendantes utilisent des statistiques sur les données pour descendre progressivement dans l'analyse.

La technologie développée par PDI utilise l'approche ascendante pour extraire des règles, ce que l'on appelle également l'induction de règles à partir de données. Cette approche consiste à généraliser des comportements locaux quand cela est pertinent.

Les avantages de cette approche sont les suivants :

- Possibilité de découvrir des situations rares (les « cas rares »)
- Capacité de prendre en compte des « cas pathologiques » dans lesquels la densité de probabilité des classes sur chaque variable n'apporte aucune information, et les approches ascendantes échouent généralement (exemple : certaines tables de décision).

- Capacité de prendre en compte un grand nombre de variables. En théorie, le temps nécessaire pour l'apprentissage varie linéairement avec le nombre de variables
- L'évaluation locale des critères par laquelle procède l'algorithme permet de s'adapter à une distribution non uniforme du bruit
- L'approche ascendante n'exige aucune évaluation globale des variables, ce qui permet l'utilisation de données sur lesquels il existe des valeurs manquantes, problème très courant dans les applications réelles
- Les interactions complexes entre variables peuvent amener les méthodes descendantes à éliminer des variables étant pertinentes dans un certain contexte. L'algorithme développé par PDI n'élimine pas de variables à priori en fonction de critères globaux, mais les élimine progressivement pour chaque règle et donc relativement à leur influence locale

C'est du fait des avantages explicités ci-dessus que nous avons choisi d'utiliser cette technique spécifique d'apprentissage sur une base de données médicale (cf. troisième partie). Nous pensons en effet que ce type d'approche convient particulièrement à la formalisation de règles dans un contexte de recueil et d'exploitation de données médicales.

DEUXIEME PARTIE :

LA NEUROFIBROMATOSE DE TYPE 1

LES PREMIERS ENSEIGNEMENTS D'UNE BASE DE DONNEES DE 87 ENFANTS ATTEINTS

I. LA NEUROFIBROMATOSE DE TYPE 1

I.1 Historique

C'est Von Recklinghausen qui, le premier, en 1882, décrit des tumeurs cutanées et profondes qu'il nomme neurofibromes. Preiser et Davenport en 1918 font l'hypothèse d'un mode de transmission autosomique dominant, qui va être confirmé par Borberg en 1951.

La distinction en deux types de Neurofibromatoses n'aura lieu qu'en 1981 par Riccardi, puis en 1988 par le National Institute of Health qui, à l'occasion d'une conférence de consensus établit les critères diagnostiques de la NF de type 1 (maladie de Recklinghausen ou neurofibromatose classique) et de la NF de type 2 (NF centrale ou acoustique).

Les critères diagnostiques des deux types de Neurofibromatoses sont en annexes (cf Tableaux n°1 et 2).

Au début des années 1990, les progrès de la génétique permettent de mettre en évidence d'abord un gène impliqué dans la NF-1 sur le chromosome 17 (17q11.2), puis un gène différent sur le chromosome 22 (22q) impliqué, lui, dans la NF-2.

C'est en 1993 qu'une base de données internationale est inaugurée à Vancouver dont le but est de collecter les caractéristiques cliniques des patients atteints de la NF-1.

I.2 Epidémiologie

L'incidence de la NF-1 à la naissance est estimée à 1/3500 naissances pour une prévalence dans la population générale d'environ 1/4500.

Il n'y a pas de prédisposition ethnique ou géographique. Le sex ratio est égal à 1.

I.3 Clinique

Il existe trois groupes de manifestations cliniques survenant au cours de la NF-1 :

- Les signes cliniques majeurs : taches café au lait, éphélides axillaires et inguinales, nodules de Lisch, neurofibromes. Ces signes, très fréquents, sont nécessaires au diagnostic de la maladie (cf. Tableau n° 1) mais n'entraînent aucun retentissement fonctionnel, à l'exception des neurofibromes cutanés.

- Les signes cliniques mineurs sont associés de manière significative à la NF-1 mais n'appartiennent pas aux critères diagnostiques. Il s'agit de la macrocéphalie, de la petite taille et de certaines anomalies morphologiques thoraciques.

- Les complications grèvent le pronostic de la maladie : certaines sont assez fréquentes (Troubles de l'apprentissage, scolioses, neurofibromes plexiformes), d'autres rares (pseudarthroses des os longs, gliome des voies optiques, anomalies endocrinologiques, épilepsie, neurofibrosarcomes, hémopathie myéloides, phéochromocytomes).

I.3.1. Les signes cliniques majeurs

a) Les taches café au lait

Les taches café au lait sont les plus fréquentes et les plus précoces des manifestations cliniques de la NF-1. Elles sont présentes dès la naissance ou apparaissent dans la première années de vie et augmentent en taille et en

nombre pendant l'enfance. Il s'agit de macules pigmentées ovalaires, de coloration brun clair, à contour régulier, d'un diamètre de 0.5 à 50 cm. L'examen histologique de ces lésions met en évidence une augmentation du contenu épidermique en mélanine sans modification du nombre des mélanocytes.

Les taches café au lait ne sont pas spécifiques de la NF-1 : 10 à 25 % des enfants sont porteurs d'une à 3 taches café au lait dans la population générale.

b) Les éphélides des plis

Les macules pigmentées retrouvées dans les plis au cours de la NF-1 sont improprement appelées éphélides. En réalité, il s'agit de petites taches café au lait, d'un diamètre inférieur à 5 mm. Leur prévalence est d'environ 80 % à l'âge de 6 ans. Les éphélides des plis apparaissent donc précocement dans la NF-1, après les taches café au lait, et constituent un signe capital dans l'enfance, permettant souvent de poser le diagnostic de la maladie chez un jeune enfant porteur de nombreuses taches café au lait.

c) Les nodules de Lisch

Les nodules de Lisch constituent un critère diagnostic de NF-1 (au moins 2 nodules dans chaque champ). Il s'agit de petits hamartomes iriens constitués d'amas de mélanocytes. De coloration brune chez l 'adulte, ils peuvent être pales chez le jeune enfant. Ces lésions sont asymptomatiques et apparaissent progressivement pendant l'enfance. Les nodules de Lisch sont exceptionnellement décrits en dehors de la NF-1 et sont quasiment pathognomoniques de cette affection.

- Les gliomes des voies optiques (les gliomes des voies optiques sont les plus fréquentes des tumeurs intracrâniennes au cours de la NF-1 et font partie des critères diagnostiques de la maladie.

- L' epilepsie

- Les autres tumeurs du système nerveux central (astrocytomes de bas grades, neurofibromes médullaires et para médullaires, méningiomes, épendymomes, schwannommes)

b) Les complications osseuses

- La pseudarthrose des os longs (très spécifique de la NF-1 il s'agit d'une lésion congénitale rare)

- Les anomalies vertébrales (accentuation de la concavité du bord postérieur de certains corps vertébraux, scolioses...)

- Les anomalies crânio faciales (dysplasie des ailes sphénoides)

c) Les complications endocrinologiques

- Puberté précoce (souvent associée à un gliome des voies optiques)
- Petite taille
- Déficits en hormone de croissance (rares et secondaires aux traitements des tumeurs cérébrales)

d) Les difficultés d'apprentissage

- Il s'agit de l'une des manifestations les plus fréquente de la NF-1, rencontrés chez 40 à 60 % des enfants. Ils sont définis par l'absence de

corrélation entre un niveau d'aptitude proche de la normale (mesuré par le quotient intellectuel) et un niveau de performance inférieure à la normale.

- Les manifestations sont très diverses allant des troubles perceptifs spatiaux et visuels entraînant des difficultés à dessiner, écrire, calculer, à se repérer dans l'espace, à des troubles de l'attention, de la mémoire immédiate, des troubles de la coordination motrice, des troubles du langage écrit et oral, ainsi que des signes d'hyperactivité.

e) Les complications vasculaires

Plusieurs localisations artérielles peuvent être décrites :

- La dysplasie des artères rénales
- La dysplasie des artères cérébrales
- La dysplasie des artères digestives

f) Les complications cutanées

- Les neurofibromes plexiformes

Il s'agit de tumeurs se développant au contact des nerfs périphériques, et présentant un aspect caractéristique en paquet de ficelle.

Sur le plan clinique, il s'agit de tuméfactions sous cutanés mal limitées, de consistance molle et irrégulière, de taille variable, recouvertes d'une hyper pilosité et d'une hyper pigmentation. Certaines lésions sont inaccessibles à l'examen clinique.

Le risque évolutif majeur est la transformation en tumeur maligne des gaines nerveuses.

- Les xanthogranulomes juvéniles sont des lésions bénignes apparaissant sous forme de papules jaunes, fermes, localisées surtout sur la tête et le cou, apparaissant dans les 2 premières années de vie pour disparaître progressivement.

g) Les complications carcinologiques

Rares chez l'enfant mais plus fréquentes que dans la population générale.

On distingue :

- Les tumeurs cérébrales de bas grade de loin les plus fréquentes (gliomes des voies optiques, astrocytomes du tronc cérébral)
- Les hémopathies myéloïdes (leucémies myéloïdes juvéniles chroniques, dysmyelopoièses, leucémies myéloïdes aiguës)
- Les tumeurs malignes des gaines nerveuses.

II. LES PREMIERS ENSEIGNEMENTS D'UNE BASE DE DONNEES DE 87 ENFANTS ATTEINTS

II.1. Objectifs et méthodologie de l'étude menée par Sébastien Barbarot.

La thèse de Sébastien Barbarot tentait d'évaluer « au cours d'une étude prospective sur 7 ans, la gravité de la maladie, la prévalence et l'incidence des signes cliniques et des complications dans une cohorte de 87 enfants porteurs d'une NF-1 ». En effet, la plupart des complications surviennent dans l'enfance et sont totalement imprévisibles.

La base de données (en annexe) provenait d'une étude monocentrique prospective multidisciplinaire et avait été constitué depuis 1992 à partir de 144 enfants examinés au Centre Nantais Neurofibromatose pour une suspicion de neurofibromatose. 87 enfants porteurs d'une NF-1 selon les critères du National Institute of Health ont été inclus dans l'étude.

Protocole de suivi

Le bilan initial comprenait pour chaque enfant :

- un examen clinique général, dermatologique, neurologique, endocrinologique, orthopédique, ophtalmologique
- une Imagerie par Résonance Magnétique (IRM) cérébrale.

Les patients étaient suivis cliniquement tous les ans en l'absence de complication intercurrente. La survenue d'une complication modifiait le rythme du suivi.

Certains examens complémentaires non systématiques pouvaient être demandés en fonction des données de l'examen clinique (dosages biologiques, radiographies standards, échographies abdominales). De même, des consultations spécialisées pouvait compléter le bilan initial ou le suivi (examen ORL, bilan phoniatrique)

La durée du suivi (*d*) était définie comme la période entre la première et la dernière consultation.

Les données enregistrées à chaque consultation étaient les suivantes :

- Description des signes cliniques majeurs :
  - Nombre et taille des taches café au lait
  - Présence et localisation des éphélides
  - Nombre de nodule de Lisch
  - Nombre de neurofibromes dermiques

- Présence d'une ou plusieurs complications. Les complications ont été définies, selon les données de la littérature, comme des événements dont la fréquence est significativement plus élevée dans la NF-1 par rapport à la population générale

Prévalence

La prévalence des signes cliniques majeurs et des complications était définie au 1er mai 1999 chez les 87 enfants porteurs d' une NF-1.

Ce groupe était défini comme le *groupe A*.

Incidence

Les taux d'incidence des complications pendant la période *d* étaient déterminés pour 2 sous groupes du groupe A :

- Les patients sans complication initiale lors de la première consultation (*sous-groupe B*, n=24).
- Les patients avec complication(s) initiale(s) lors de la première consultation (*sous-groupe C*, n=30).

Les patients non suivis ou suivis pendant une durée *d* inférieure à 1 an constituaient le *sous groupe D* (n=33).

Evaluation de la gravité de la maladie

La gravité de la maladie était évaluée grâce aux grades de sévérité établis par Riccardi

Grades de sévérité de la maladie

- **Grade 1:** taches café au lait seules, ou avec un petit nombre de neurofibromes cutanés sans conséquence cosmétique ou fonctionnelle.
- **Grade 2**: neurofibromes cutanés en grand nombre et/ou entraînant des conséquences cosmétiques ou fonctionnelles limitées. Lésions osseuses asymptomatiques.

- **Grade 3**: neurofibromes cutanés en grand nombre, neurofibromes viscéraux, manifestations orthopédiques (pseudarthrose, scoliose), épilepsie bien contrôlée, hypertrophie locale modérée.

- **Grade 4**: atteinte sévère avec retentissement majeur incluant: tumeurs intracrâniennes et spinales, schwannomes malins, neurofibrosarcomes, phéochromocytomes, épilepsies non contrôlées, retard mental, hydrocéphalies, et hypertrophie progressive ou diffuse.

II.2. Résultats et hypothèses

Les résultats de l'étude de Sébastien Barbarot sont les suivants :

« Il s'agit de la deuxième étude évaluant l'incidence des complications de la NF-1 dans une population pédiatrique de 87 patients. Plusieurs points doivent être soulignés :

- Nous confirmons la séquence d'apparition des signes majeurs de la maladie et leur intérêt dans le diagnostic précoce.
- Nous mettons en évidence pour la première fois une incidence totale de 13 complications pour 100 patients année.
- Nous montrons que la majorité des complications apparaissant dans l'enfance sont les difficultés d'apprentissage. Ces troubles doivent être considérés comme des complications à part entière en raison de leur retentissement.

- Nous confirmons la nécessité d' un suivi clinique régulier multidisciplinaire pendant l'enfance car les complications sont fréquentes, imprévisibles et parfois évolutives (hydrocéphalies, gliomes des voies optiques, anomalies endocrinologiques, neurofibromes plexiformes).
- Nous confirmons que seuls 30 % des enfants ont un retentissement mineur de la maladie.

Nous suggérons de réévaluer l'intérêt de l'IRM cérébrale précoce systématique pour le dépistage et le suivi des patients " à risque " de complications (hydrocéphalies, GVO).»

- Face à ces premiers résultats il nous a donc semblé opportun d'appliquer les algorithmes d'apprentissage non paramétriques à cette base de données afin de tenter de répondre à une série d'hypothèses : Un enfant atteint de NF-1 et de profil uniquement cutané, présentera t'il moins de complications à type de gliomes ou de troubles de l'apprentissage qu'en enfant ayant un profil mixte ? Existe-t-il des règles de corrélations entre la présence de tel ou tel symptôme et l'existence d'une pseudarthrose ? Existe-t-il des règles de corrélations entre différents symptômes de la maladie et le sexe de l'enfant atteint ?

TROISIEME PARTIE :

UN ALGORITHME D'APPRENTISSAGE NON-PARAMETRIQUE A L'EPREUVE DE LA BASE DE DONNEES DES 87 ENFANTS ATTEINTS DE NEUROFIBROMATOSE

Dans tout travail de datamining, et quelque soit la méthode algorithmique utilisée, il est impératif de « reformater » la base de donnée de laquelle on entend extraire de la connaissance (I).

Cela signifie qu'en préalable il nous faut nous interroger sur la compatibilité du format de cette base de données avec le logiciel utilisé.

La notion de « format » est entendu au sens large : c'est-à-dire d'une part le format informatique (base de données excel©, access©, oracle©, sybase© ...), et d'une autre la façon dont les données ont été recueillies en regard des connaissances que nous cherchons à extraire.

Ce travail préalable effectué, nous pourrons alors tenter d'extraire des règles sur la neurofibromatose, grâce au logiciel, à partir d'hypothèses élaborées par l'expert (II).

Trois hypothèses ont été retenues et sont formulées ainsi :

- Les enfants ayant des complications de types difficultés d'apprentissage ont-ils un profil particulier par rapport aux enfants ne présentant pas de telles difficultés d'apprentissage ?
- Existe-t-il des règles sur la pseudarthrose ?
- Existe-t-il des règles sur le sexe des enfants atteints de neurofibromatose ?

I. REFORMATAGE DE LA BASE DE DONNEES

La base de données utilisée par Sébastien Barbarot se présente sous la forme d'un tableau Excel©. Les lignes correspondent à chacun des 87 patients, les colonnes correspondent aux caractéristiques relevés sur chacun de ces patients (le sexe, l'âge, le nombre de taches café au lait, la présence ou non d'un gliome des voies optiques, d'une hyperdensité en T2 à l'IRM etc.).

Un extrait de cette base se présente ainsi de la façon suivante :

N°	Sexe	Age	B	C	D	TCL> 6	TCL<6	Eph	NF0	NF0-10	NF<100	NF>100	Lisch	Diff Appr	NF plex	GVO	Dysplos	Hyper T2	Autres	Signes Clin
13		0,5	5	1	7	12	1	10	11	2			4	5	1	1		6	2	
59	F	2			X	X		X	X										angiome plan joue	
39	F	2,1			X	X			X											
50	M	3,4			X	X		X	X				X					X		
58	F	3,5			X	X		X	X											
81	M	3,5			X	X		X	X											
2	F	3,8	X			X		X	X					X	X			X	Céphalées ss Zaditen	
10	M	4,1	X				X		X									X		
74	F	4,5	X			X			X									X		
21	M	4,6			X	X		X		X			X	X				X		
23	M	4,6	X			X		X		X										J/K
36	F	4,9	X			X		X	X					X						
35	M	5			X	X		X	X				X	X						
63	M	5		C1		X		X	X				X	X		Xstab		X		
27		5,1	11	5	11	26	1	26	14	13			12	9	3	4	4	20	9	
69	F	5,3			X	X		X	X				X(1)					X		
49	M	5,5	X			X		X	X											
86	F	5,9	X			X		X		X 1opr								X	XGJ?	I J/K
40	M	6		C1		X		X		X			X				pseudopr	X		I J/
66	F	6,7	X			X		X	X									X		
15	F	6,8			X		X	X	X					XX						
83	F	6,8	X			X		X		X			X	X	Xcervical			X	sténose acq Sylvius	I J/K N
14	M	7			X	X		X		X			X					X		
9	F	7,1	X			X		X	X				X	X				X		N
68	F	7,1			X	X		X	X									X		
84	F	7,7			X	X		X	X									X		
16	F	8			X	X		X	X					XXX			Scolose	X	Scolose	
25	M	8		C1		X		X	X				X			Xstab		Xdiminué		N
47	M	8,4	X			X		X		X									LAL +retard de croiss	J/K
43	F	8,7		C1		X		X	X				X(3)			X TT		X	sténose acq Sylvius	I N
18	F	9			X	X		X	X				X	X		X		X	Avce.staturale	
55	F	9			X	X		X	X				X	X			inegmi	X	mégagrde citerne	
73	F	9			X	X		X		X			X					X		
87	M	9	X			X		X		X			X			X		X		I J/K N
67	F	9,1	X			X		X		X					X axill G			X		J/K
4	F	9,2	X			X		X		X			X	X				X	Céphalées	J/K N
37	M	9,4			X	X		X		X										
44	F	9,4	X			X		X	X											I

Pour avoir manipulé un nombre assez important de bases de données de type « médicale », nous pouvons dire de prime abord qu'il s'agit ici d'une base de données de bonne qualité globale. C'est-à-dire, et sans préjuger de la véracité de l'information (par exemple il aurait pu y avoir une erreur de recopiage faussant tous les résultats) que les données sont relativement complètes et homogènes.

Nous noterons cependant que dans un certain nombre de cas les informations sont de nature ou de degré différents. Par exemple dans la colonne NF-10 (qui signifie nombre de neurofibromes inférieur à 10), nous aurons la plupart du temps une croix

(X) qui signifie que le patient est porteur de moins de 10 localisations de neurofibromes, mais de temps en temps ce n'est plus une croix (X) mais le nombre exact de localisation est précisé (2 par exemple). Parfois même il est rajouté un renseignement qui avait peut être semblé intéressant à l'opérateur mais qui empêchera tout algorithme de fonctionner correctement (par exemple une case comprend l'élément X1opr qui semble vouloir dire qu'il existe une seule localisation de neurofibrome qui a subi une opération d'exérèse). L'adage populaire « abondance de biens ne nuit pas » est dans le cas du datamining particulièrement faux car le logiciel ne va pas pouvoir interpréter cette information supplémentaire. Il va donc en résulter, dans le meilleur des cas, une perte d'information brute, par défaut d'interprétation de la donnée.

Plus dangereux encore est l'incohérence de l'information. Toujours à propos de la colonne NF-10 on peut noter le nombre 13 dans une des cases. Le patient concerné présente t'il effectivement moins de 10 localisations (par exemple 1 ou 3) ou bien présente t'il 13 localisations auquel cas une croix (X) devrait se trouver dans la colonne NF-100 (nombre de localisations inférieure à 100) ?

Dans ce cas notre travail sera de tenter de récupérer l'information à sa source. Si le doute est impossible à lever il nous faudra effacer cette information pour éviter de fausser l'extraction de la connaissance par le logiciel.

Un autre travail de préparation de la base de données va être le fait de s'interroger sur la logique de l'information. Prenons le cas d'un patient présentant 8 localisations de neurofibromes. Faut il considérer une croix (X) et une seule dans la colonne NF-10, ou faut il considérer une croix (X) dans la colonne NF-10 et une croix (X) dans la colonne NF-100 ? Les deux réponses sont vraies et pour savoir comment trancher il nous faut nous interroger sur la façon dont l'ordinateur traite l'information d'une part,

sur l'objectif de la réponse que nous attendons du logiciel d'autre part (objectif de segmentation de la population (une seule croix) objectif d'exhaustivité de la réponse (deux croix)).

Nous devons reconnaître qu'il s'agit de la partie de la préparation des données la plus délicate et qu'en l'absence de méthodologies validées en l'espèce, nous n'aurons parfois d'autres choix que d'utiliser notre intuition.

De façon plus systématique nous allons souvent être amené à synthétiser l'ensemble des informations se rapportant à un critère dans une seule colonne. Par exemple les 4 colonnes initiales de la base de données (NF-0, NF-10, NF-100, NF>100) vont être remplacées par une unique colonne « Nombre de Neurofibromes » contenant dans chacune des cases l'une des 4 valeurs suivantes : pas, <10, <100, >100.

Au contraire il nous faut dans d'autres cas subdiviser une seule colonne en plusieurs. Par exemple la base de données initiale comporte une colonne appelé « dysplasie osseuse » dans laquelle on retrouve des éléments de signification très différentes (inégalité des membres inférieurs et pseudarthrose). Il s'agit d'un cas typique où seul l'expert peut nous orienter. En l'occurrence, des interrogations subsistant par rapport au liens existant entre la pseudarthrose et la neurofibromatose, il est décidé de créer une colonne unique « pseudarthrose » dans laquelle les valeurs pourront être « non » ou « oui » signifiant présence ou non d'une pseudarthrose.

La base de données finale (extrait) se présente donc ainsi :

Sexe	Age	Complications lors 1ère cs	Classe C	Nb de TCL	Ephélide	Nb de Neurofibrome	Lisch	Difficulté Apprentissage	retentissement NF plex	Gliome VO	pseudarthrose	Hyper T2	Tr. IRM ou autres	Signes Clin	Complic. par diff d'appr.
F	3,8	B		> 6	oui	pas	non	1	faible	non	non	oui	tr. Cerebr.		oui
F	9,2	B		> 6	oui	<10	oui	1	pas de NF	non	non	oui	tr. Cerebr.	J/K N	oui
F	7,1	B		> 6	oui	pas	oui	1	pas de NF	non	non	oui	sans	N	oui
M	4,1	B		< 6	non	pas	non	0	pas de NF	non	non	oui	sans		non
M	4,6	B		> 6	oui	<10	non	0	pas de NF	non	non	non	sans	J/K	non
F	11,6	B		> 6	oui	<10	non	1	pas de NF	non	non	non	sans		oui
M	13,5	B		> 6	oui	<10	oui	1	pas de NF	non	non	oui	sans	I N	oui
F	4,9	B		> 6	oui	pas	non	1	pas de NF	non	non	non	sans		oui
F	10,8	B		> 6	oui	pas	non	0	pas de NF	non	non	oui	tr. Cerebr.	N	non
F	9,4	B		> 6	oui	pas	non	0	pas de NF	non	non	non	sans	I	non
M	8,4	B		> 6	oui	<10	non	0	pas de NF	non	non	non	tr. Cerebr.	J/K	non
M	5,5	B		> 6	oui	pas	non	0	pas de NF	non	non	non	sans		non
M	14,1	B		> 6	oui	<10	non	0	pas de NF	non	non	oui	sans	J/K	non
F	12,4	B		> 6	oui	<100	oui	0	pas de NF	non	non	oui	sans	K/L	non
F	6,7	B		> 6	oui	pas	non	0	pas de NF	non	non	oui	sans		non
F	9,1	B		> 6	oui	<10	non	0	faible	non	non	oui	sans	J/K	non
F	11	B		> 6	oui	pas	oui	0	pas de NF	non	non	oui	sans		non
M	15	B		> 6	oui	<100	non	0	pas de NF	non	non	oui	sans	K/L	non
F	4,5	B		> 6	non	pas	non	0	pas de NF	non	non	oui	sans		non
F	17,7	B		> 6	oui	pas	non	0	pas de NF	non	non	oui	sans		non
F	9,8	B		> 6	oui	<10	non	0	pas de NF	non	non	oui	sans	J/K	non
F	6,8	B		> 6	oui	<10	oui	1	faible	non	non	oui	tr. IRM	I J/K N	oui
F	5,9	B		> 6	oui	<10	non	0	pas de NF	non	non	oui	tr. Cerebr.	I J/K	non
M	9	B		> 6	oui	<10	oui	0	pas de NF	oui	non	oui	sans	I J/K N	non
M	13,5	C	C1	> 6	oui	<10	non	2	pas de NF	non	non	non	sans	J/K	oui
F	10,1	C	C1	> 6	oui	pas	oui	1	pas de NF	non	non	oui	sans	I	non
F	16	C	C2	> 6	oui	<10	oui	0	important	non	non	oui	sans	J/K	non
F	16,5	C	C1	> 6	oui	<100	oui	2	pas de NF	non	non	non	tr. IRM	K/L	oui
F	10	C	C2	> 6	non	<10	non	0	important	non	non	non	sans	J/K	non
M	17	C	C2	> 6	oui	<10	non	1	important	non	non	oui	sans	J/K	non
M	8	C	C1	> 6	oui	pas	oui	0	pas de NF	oui	non	oui	sans	N	non
F	13,6	C	C2	> 6	oui	pas	oui	2	pas de NF	non	non	oui	sans	N+X	non
M	9,6	C	C1	> 6	oui	pas	non	2	pas de NF	non	non	oui	tr. IRM		oui
M	11,7	C	C2	> 6	oui	pas	oui	2	pas de NF	non	non	non	sans	I N	non
F	8,7	C	C1	> 6	oui	pas	oui	0	pas de NF	oui	non	oui	tr. IRM	I N	non
F	12	C	C1	> 6	non	<10	oui	1	pas de NF	non	non	oui	tr. Cerebr.		non
M	13	C	C3	> 6	oui	pas	non	2	pas de NF	non	non	non	tr. IRM		non
M	12,1	C	C1	> 6	non	<10	oui	2	pas de NF	non	non	oui	tr. Cerebr.	J/K	oui
M	18	C	C2	> 6	oui	<100	oui	2	pas de NF	non	non	non	sans		non
M	16,5	C	C1	> 6	oui	<10	oui	2	faible	non	non	oui	sans		oui
M	5	C	C1	> 6	oui	pas	oui	1	pas de NF	oui	non	oui	sans		oui
M	15	C	C2	> 6	oui	<10	oui	2	faible	oui	non	non	sans	J/K N	oui
M	15,5	C	C1	> 6	oui	pas	oui	2	pas de NF	non	non	non	sans		non

A l'issue de cette « datapreparation » si nous devions attirer l'attention sur un seul point susceptible d'améliorer énormément la qualité d'une base de données médicale sans coûter trop d'effort à l'opérateur qui recueille et intègre les données dans la base, ce serait sur « l'homogénéisation » des données.

L'opérateur doit se tenir strictement à la contrainte de remplissage (pas de valeurs numériques s'il s'agit d'un booléen, pas de commentaires supplémentaires ou bien dans une autre colonne distincte...)

Evidement ces contraintes se doivent d'avoir été clairement définies en amont.

Ces contraintes se doivent d'apparaître en aval le plus explicitement possible sur le formulaire de remplissage des données puisque l'opérateur ne sera quasiment jamais le même le temps de la vie de la base de données.

II. EXTRACTION DE REGLES SUR LA FIBROMATOSE GRACE AU LOGICIEL, À PARTIR D'HYPOTHESES ELABOREES PAR L'EXPERT

II.1 Les enfants ayant des complications de types difficultés d'apprentissage ont-ils un profil particulier par rapport aux enfants ne présentant pas de telles difficultés d'apprentissage ?

Pour tenter de répondre à cette question il nous faut effectuer une data preparation supplémentaire.

En effet nous allons considérer deux populations : celle comprenant les enfants sans troubles de l'apprentissage et celle comprenant les enfants ayant des troubles de l'apprentissage.

Nous allons donc fusionner deux colonnes de la base de données initiale : la colonne « difficultés d'apprentissage » cotée de 0 (pas de difficultés d'apprentissage) à 3 (difficultés d'apprentissage importantes) et la colonne « complications à type de troubles de l'apprentissage ». Cette nouvelle colonne se nommera « difficultés d'apprentissage » et sera remplie de façon binaire :

- soit l'enfant a eu des difficultés d'apprentissage initiales et/ou des complications à type de troubles de l'apprentissage, auquel cas la donnée sera « oui » (groupe avec difficultés d'apprentissage)
- soit l'enfant n'a jamais eu ni de difficultés d'apprentissage initiale ni de complications à type de troubles de l'apprentissage, auquel cas la données sera « non » (groupe sans difficultés d'apprentissage).

La première série de tests effectués ne nous donnent pas de règles pertinentes c'est pourquoi nous effectuons un nouveau travail de data preparation : le groupe sans « difficultés d'apprentissage » comprendra également les enfants ayant été considérés comme présentant des difficultés d'apprentissage faibles d'une part (cotation 1 et non plus seulement 0) mais ne présentant évidemment toujours pas de « complications à type de troubles de l'apprentissage ».

La base se présente alors ainsi (extrait) :

Sexe	Age	Comp	Class	Nb de TCL	Ephéride	Nb de Neurofibro	Lisch	Difficulté d'apprentissage	retentissement NF plex	Gliome VO	pseudarthros	scoliose	Ineg mb inf	Hyper T2
F	15,7	D		> 6	oui	<10	oui	non	faible	non	non	oui		non
M	17,9	D		> 6	oui	<100	oui	non	faible	oui	oui	non	non	non
M	19,1	C	C3	> 6	oui	<10	oui	non	faible	non	non	non	non	non
F	9,1	B		> 6	oui	<10	non	non	faible	non	non	non	non	oui
F	14,5	D		> 6	oui	pas	oui	non	faible	non	oui	non	non	oui
M	14,1	B		> 6	oui	<10	non	non	pas de NF	non	non	non	non	oui
M	14,1	D		> 6	oui	<10	oui	non	pas de NF	non	non	non	non	oui
M	14,1	C	C1	> 6	oui	>100	oui	non	pas de NF	non	non	non	oui	oui
M	15	B		> 6	oui	<100	non	non	pas de NF	non	non	non	non	oui
F	17,7	B		> 6	oui	pas	non	non	pas de NF	non	non	non	non	oui
F	19	C	C3	> 6	oui	<10	oui	non	faible	non	non	non	oui	oui
M	17	C	C2	> 6	oui	<10	non	non	important	non	non	non	non	oui
M	5	D		> 6	oui	pas	oui	non	pas de NF	non	non	non	non	non
M	12	D		> 6	oui	<10	oui	non	pas de NF	non	non	non	non	non
M	17	D		> 6	oui	pas	oui	non	pas de NF	non	non	non	non	non
M	17,1	D		< 6	oui	pas	non	non	pas de NF	non	non	non	non	non
M	4,6	D		> 6	oui	<10	oui	non	pas de NF	non	non	non	non	oui
F	9	D		> 6	oui	pas	oui	non	pas de NF	oui	non	non	non	oui
F	9	D		> 6	oui	pas	oui	non	pas de NF	non	non	non	oui	oui
F	10,1	C	C1	> 6	oui	pas	oui	non	pas de NF	non	non	non	non	oui
F	12	C	C1	> 6	non	<10	oui	non	pas de NF	non	non	non	non	oui
M	14,6	C	C2	> 6	oui	pas	non	non	pas de NF	non	non	non	oui	oui
M	18	D		> 6	oui	<10	oui	non	pas de NF	non	non	non	non	oui
F	6,8	D		< 6	oui	pas	non	oui	pas de NF	non	non	non	non	non
M	11,7	C	C2	> 6	oui	pas	oui	oui	pas de NF	non	non	non	non	non
M	13	C	C3	> 6	oui	pas	non	oui	pas de NF	non	non	non	non	non
M	15,5	C	C1	> 6	oui	pas	oui	oui	pas de NF	non	non	non	non	non
M	18	C	C2	> 6	oui	<100	oui	oui	pas de NF	non	non	non	non	non
M	12	D		> 6	oui	pas	non	oui	pas de NF	non	non	non	non	oui
M	13,5	C	C3	> 6	oui	pas	oui	oui	pas de NF	non	non	non	oui	oui
F	13,6	C	C2	> 6	oui	pas	oui	oui	pas de NF	non	non	non	non	oui
M	16,1	D		> 6	oui	<100	oui	oui	pas de NF	non	non	non	oui	non
M	9,5	D		> 6	oui	<10	non	oui	pas de NF	non	non	non	non	non
F	8	D		> 6	oui	pas	non	oui	pas de NF	non	non	oui	non	oui
M	6	C	C1	> 6	oui	<10	oui	oui	pas de NF	non	oui	non	non	oui
F	3,8	B		> 6	oui	pas	non	oui	faible	non	non	non	non	oui
F	6,8	B		> 6	oui	<10	oui	oui	faible	non	non	non	non	oui
M	14	C	C3	> 6	oui	<10	oui	oui	faible	non	oui	non	non	oui
F	4,9	B		> 6	oui	pas	non	oui	pas de NF	non	non	non	non	non
F	11,6	B		> 6	oui	<10	non	oui	pas de NF	non	non	non	non	non
M	5	C	C1	> 6	oui	pas	oui	oui	pas de NF	oui	non	non	non	oui
F	7,1	B		> 6	oui	pas	oui	oui	pas de NF	non	non	non	non	oui
F	9,2	B		> 6	oui	<10	oui	oui	pas de NF	non	non	non	non	oui
M	13,5	B		> 6	oui	<10	oui	oui	pas de NF	non	non	non	non	oui

Nous noterons que 57 enfants appartiennent au groupe « sans difficultés d'apprentissage » et 30 enfants appartiennent au groupe « avec difficultés d'apprentissage » soit environ 2 fois moins.

Il s'agit ensuite de paramétrer le logiciel afin que le résultat des règles qu'il va extraire de la base de données soit l'existence ou non de difficultés d'apprentissage. Les règles, si elles existent, se présenteront ainsi sous la forme :

Si A alors « difficultés d'apprentissage » = oui

Ou bien Si B alors « difficultés d'apprentissage » = non

La règle la plus pure (c'est-à-dire ne contenant pas de contre-exemple) extraite à l'aide de l'algorithme porte sur l'absence de « difficultés d'apprentissage » (ce qui n'est pas étonnant étant donné que la population « sans difficultés d'apprentissage » est deux fois plus nombreuse) et s'inscrit ainsi :

Rule 1: non

Purity: non: 100%, oui: 0%

Size: samples: 6, volume: 29%

If	Age	in [2, 12]
and	Ephélide	= non
then	Difficulté d'apprentissage	= non

Cette règle nous indique donc que :

Tout enfant dans la base de données d'âge inférieur à 12 ans (il n'y a pas d'enfant d'âge inférieur à 2 ans dans la base) et ne présentant pas d'éphélides n'aura pas de difficultés d'apprentissage.

Evidemment la population n'est que de 6 individus et il est hors de question d'en faire un règle statistiquement significative, cependant il serait nécessaire d'étudier la relation entre l'absence d'éphélide chez des enfants avant la puberté environ et la survenue de difficultés d'apprentissage, à l'aide d'une base de données plus significative.

Il s'agit ici d'une utilisation du logiciel que l'on pourrait qualifier de « prospective » avec l'émergence de pistes, non significative statistiquement mais susceptibles d'orienter des recherches plus approfondis.

II.2 Existe-t-il des règles sur la pseudarthrose ?

Nous nous intéressons à présent à une autre hypothèse ouverte : existe-t-il dans la base de données des corrélations entre la présence ou non de tel ou tel symptômes et l'existence d'une pseudarthrose ?

Il n'existe pas aujourd'hui d'études démontrant de façon probante l'existence d'une telle relation.

Il s'agit de laisser l'algorithme chercher les règles d'associations les plus pertinentes entre la présence ou non d'une pseudarthrose et les différents attributs renseignés dans la base de données.

Que veut dire « les règles les plus pertinentes » pour le logiciel ?

Cela signifie que vont être extraites les règles regroupant le plus d'individus possibles (l'échantillon aura donc plus de chance d'être représentatif au sens statistique) et les règles les plus « pures » (c'est-à-dire les règles qui ne génèrent pas de contre-exemple).

L'algorithme fait un compromis entre la pureté et le volume de l'échantillon en fonction des indications de paramétrages que l'on lui donne.

En l'occurrence nous réglons le paramètre de pureté à 100% afin que le travail de l'algorithme soit de trouver les règles d'associations regroupant le plus d'individus possible.

Les résultats obtenus sont les suivants :

- 7 règles sont extraites de la base
- toutes les règles concernent l'absence de pseudarthrose et donc aucune la présence de pseudarthrose (ce qui n'est pas étonnant puisqu'il n'existe que 4 cas de pseudarthrose sur les 78 cas)
- le volume de la population agrégée va de 22 individus (règle n°1) à 58 individus (règle n°7)
- la complexité des règles est d'ordre 1, 2 ou 3[2].

Voici les règles en détail :

Rule 1: non

Purity: non: **100%**, oui: 0%

Size: samples: **22**,

If	Age	in [6.7, 19.1]
and	Nb de Neurofibrome	= <10
and	Tr. IRM ou autres cerebral	= sans
then	pseudarthrose	= non

Rule 2: non

Purity: non: **100%**, oui: 0%

Size: samples: **32**, volume: **36%**

If	Sexe	= F
and	Age	in [2, 14.2]
then	pseudarthrose	= non

Rule 3: non

Purity: non: **100%**, oui: 0%

Size: samples: **34**,

If	Sexe	= M
and	Gliome VO	= non
and	Complic. par diff d'appr.	= non
then	Pseudarthrose	= non

Rule 4: non

Purity: non: **100%**, oui: 0%

[2] Si A alors X (la complexité est d'ordre 1), si A et B alors X (la complexité est d'ordre 2), si A et B et C alors X (la complexité est d'ordre 3).

Size: samples: **37**,

If Lisch = non
then pseudarthrose = non

Rule 5: non

Purity: non: **100%**, oui: 0%

Size: samples: **42**,

If Nb de Neurofibrome = pas
and retentissement NF plex = pas de NF
then Pseudarthrose = non

Rule 6: non

Purity: non: **100%**, oui: 0%

Size: samples: **55**,

If Age in [6.7, 19.1]
and retentissement NF plex = pas de NF
then Pseudarthrose = non

Rule 7: non

Purity: non: **100%**, oui: 0%

Size: samples: **58**,

If retentissement NF plex = pas de NF
and Complic. par diff d'appr. = non
then Pseudarthrose = non

Discussion :

Chaque règle comprend un nombre d'individus significatifs (jusqu'à 58 individus sur 78 au total pour la règle n°7) et la pureté de chaque règle est de 100%. Cependant le simple fait que seul soit recensé 4 cas de pseudarthrose entraîne l'impossibilité d'exploiter ces règles.

En effet si l'on prend la règle n°4 qui est très simple à comprendre puisqu'elle est de complexité 1 :

If Lisch = non
then pseudarthrose = non

Cela signifie simplement que les 4 cas de pseudarthrose sont associés à la présence de nodules de Lisch. On ne pourra pas dire que si on trouve des nodules de Lisch chez un patient, on trouvera une pseudarthrose évidemment. On pourrait dire que

dès qu'il existe une pseudarthrose chez un patient, celui-ci sera porteur de nodules de Lisch. Cependant, là encore, il n'est pas besoin d'aller plus loin dans les statistiques pour comprendre que le nombre trop faible de l'échantillon ne permet pas de conclure de la sorte.

Le nombre d'individus inclus dans cette règle n°4 est de 37 (sur un total rappelons-le de 87) ce qui est dans des proportions cohérentes avec le ratio usuellement retrouvé de non porteurs de nodules de Lisch sur les porteurs. On ne doit donc pas crier victoire trop vite en considérant que la règle couvre un nombre de profils significatifs dans la base de données.

Le raisonnement suivi peut s'appliquer aux autres règles de complexités supérieures. Et nous sommes donc dans l'impossibilité de conclure à autre chose qu'il n'existe aucune règle pertinente dans cette base de données, permettant de corréler la présence ou non de pseudarthrose avec les autres symptômes de la maladie.

II.3. Existe t il des règles sur le sexe ?

L'exemple suivant a été choisi pour illustrer à la fois la puissance de tels outils et leurs limites souvent inhérentes au faible volume de la base de données.

Le mode opératoire suivi est le même que celui de l'exemple précédent. Il est demandé au logiciel de trouver les meilleures règles permettant de corréler le sexe de l'enfant porteur de NF-1 avec les symptômes de la maladie.

La pureté du logiciel est paramétrée sur 100% pour éviter d'avoir des contre-exemples dans les règles.

Les résultats obtenus sont les suivants :

- 7 règles sont extraites de la base
- 6 règles concernent la prédiction sur le sexe masculin et 1 règle concerne la prédiction sur le sexe féminin
- le volume de la population agrégée va de 5 à 6 individus
- la complexité des règles est d'ordre 2,3,4 et 5

Voici les règles en détail :

Rule 2: M

Purity: M: 100%, F: 0%

Size: samples: 5,

If	Nb de TCL	= > 6
and	Difficulté Apprent.	= 2
and	Hyper T2	= non
and	Tr. IRM ou autres cerebral	= sans
then	Sexe	= M

Rule 3: M

Purity: M: 100%, F: 0%

Size: samples: 5,

If	Ephélide	= oui
and	Lisch	= oui
and	Gliome VO	= non
and	Tr. IRM ou autres cerebral	= tr. Cerebr.
and	Complic. par diff d'appr.	= non
then	Sexe	= M

Rule 4: M

Purity: M: 100%, F: 0%

Size: samples: 5,

If	Nb de TCL	= > 6
and	Lisch	= non
and	Difficulté Apprent.	= 2
then	Sexe	= M

- **Rule 1: F**

Purity: F: **100%**, M: 0%

Size: samples: **6**,

If	Nb de TCL	= > 6
and	Nb de Neurofibrome	= pas
and	Lisch	= non
and	Difficulté Apprent.	= 0
and	Hyper T2	= oui
then	Sexe	= F

- **Rule 5: M**

Purity: M: **100%**, F: 0%

Size: samples: **6**,

If	Nb de Neurofibrome	= <10
and	Difficulté Apprent.	= 2
then	Sexe	= M

- **Rule 6: M**

Purity: M: **100%**, F: 0%

Size: samples: **6**,

If	Ephélide	= oui
and	Nb de Neurofibrome	= <10
and	scoliose	= non
and	Hyper T2	= non
and	Complic. par diff d'appr.	= non
then	Sexe	= M

- **Rule 7: M**

Purity: M: **100%**, F: 0%

Size: samples: **6**,

If	Ephélide	= oui
and	Nb de Neurofibrome	= <10
and	ineg mb inf	= non
and	Hyper T2	= non
and	Complic. par diff d'appr.	= non
then	Sexe	= M

Discussion :

Dans cette hypothèse on se trouve en problème différent de l'exemple précédent : il s'agit du nombre très faible d'individus (5 et 6 en fonction des cas) agrégés par chacune des règles.

Le logiciel étant paramétré pour trouver les règles comportant le plus grand nombre d'individus on peut être certain qu'il n'existe, dans cette base de données, aucune règle prédictive du sexe agrégeant plus de 6 individus et restant pure à 100%.

Il est déraisonnable de se forger une conviction sur des règles de si faible volume.

On peut cependant considérer ces différentes règles comme point de départ pour une recherche plus poussée. Par exemple recueillir d'avantages de données, notamment celles qui semblent apparaître le plus souvent dans les différentes règles et pouvoir ainsi réutiliser le logiciel avec ces nouvelles données et tenter d'obtenir des volumes d'individus agrégés plus significatifs.

Le logiciel reste donc dans ce cas un outil intéressant d'exploration rapide des données, qui permet d'orienter dans un second temps et de rationaliser la collecte future d'informations.

ANNEXES

Tableau n° 1 : Critères diagnostiques de la NF-1

NF-1
Au moins 2 des critères suivants présents :
1. Avant 12 ans : au moins 6 taches café au lait d'un diamètre supérieur à 5 mm. Après 12 ans : au moins 6 taches café au lait d'un diamètre supérieur à 15 mm.
1. Au moins 2 neurofibromes ou 1 neurofibrome plexiforme
2. Pseudo-éphélides axillaires ou inguinales
3. Gliome des voies optiques
4. Au moins 2 hamartomes iriens (nodules de Lisch)
5. 1 lésion osseuse spécifique (dysplasie sphénoidale, amincissement de la corticale d' un os long avec ou sans pseudarthrose)
6. Antécédent familiaux directs de NF- 1

Tableau n° 2: Critères diagnostiques de la NF-2.

NF-2
Schwannomes vestibulaires bilatéraux (visualisés par IRM) Ou :
Antécédent familial direct de NF-2 et un des critères suivants :
a) Schwannome vestibulaire unilatéral diagnostiqué avant 30 ans.
b) Deux des affections suivantes : méningiome, gliome, schwannome, cataracte sous capsulaire postérieure juvénile.

Tableau n° 3: Base de données complète.

N°	Sexe	Age B	C	D	TCL> 6	TCL<6	Eph	NF0	NF0-10	NF<100	NF>100	Lisch	Diff Appr	NF plex	GVO	Dyspl os	Hyper T2	Autres
1	M	15	C1		X		X		X				XX					
2	F	3,8 X			X		X	X					X	X			X	Céphalées ss Zadiden
3	M	11,5	C1		X		X	X				X				Inégalité MI		
4	F	9,2 X			X		X		X			X	X				X	Céphalées
5	M	13,1		X	X		X	X				X				Inégalité MI	X	
6	M	14,2		X	X		X			X		X						Chir NF
7	M	12		X	X		X		X			X	X					
8	F	10,1	C1		X		X	X				X	X				X	
9	F	7,1 X			X		X	X				X	X				X	
10	M	4,1 X				X		X									X	
11	F	16	C2		X		X		X			X		XX			X	
12	F	18,5	C1		X		X			X		X	XX					Tum hemisph.Gstable
13	M	14,1		X	X		X		X			X					X	X.G.J.
14	M	7		X	X		X		X			X					X	
15	F	6,8		X		X	X	X					XX					
16	F	8		X	X		X	X					XXX			Scoliose	X	Scoliose
17	M	17		X	X		X	X				X	X					Malform.fosse post.
18	F	9		X	X		X	X				X	X		X		X	Avce staturale
19	M	9,5		X	X		X		X				XXX			X		
20	F	10	C2		X				X					XXopéré				
21	M	4,6		X	X		X		X			X	X				X	
22	M	17	C2		X		X		X				X	XX			X	
23	M	4,6 X			X		X		X									
24	F	11,6 X			X		X		X				X					
25	M	8	C1		X		X	X				X			X	Xstab		Xdiminué
26	F	13,6	C2		X		X	X				X(+1)	XX				X	
27	F	15,7		X	X		X		X			X		X		scol TT		
28	M	9,6	C1		X		X	X					XX				X	hydrocéph.op+XGJ
29	M	11,7	C2		X		X	X				X(1)	XX					
30	M	13,5 X			X		X		X			X	X					Xdiminué
31	M	18		X	X		X		X			X	X				X	
32	M	15,5	C1		X		X		X			X(+1)	XX			inégmiop		malform.fosse post
33	M	17,1		X		X	X	X					X					XGJ choroide
34	M	12,7	C3			X	X	X					XX(-1X)			scoliose		mégagrde citerne
35	M	5		X	X		X	X				X	X					
36	F	4,9 X			X		X	X					X					
37	M	9,4		X	X		X		X									
38	M	14,1	C1		X		X				X	X				inégmi	X	Eph intra abdo
39	F	2,1		X	X			X										
40	M	6	C1		X		X		X			X				pseudopr	X	
41	M	19	C3		X		X		X				XX	XXocc		scol TT		retard de croiss
42	F	10,8 X			X		X	X									X	saillie stern. amélio
43	F	8,7	C1		X		X	X				X(3)			X	TT	X	sténose acq Sylvius
44	F	9,4 X			X		X	X										
45	M	14,6	C2		X		X	X					X			inégmi	X	nystag. +strabisme
46	F	12	C1		X				X			X	X				X	puberté précoce
47	M	8,4 X			X		X		X									LAL +retard de croiss
48	M	13	C3		X		X	X					XX					hydrocéph retardcrois
49	M	5,5 X			X		X	X										
50	M	3,4		X	X		X	X				X					X	
51	M	13,5	C3		X		X	X				X	XX			inégmiop	X	aspect tumpédoncule
52	M	12,1	C1		X				Xcostal			X	XX				X	pectus exc
53	F	10,5		X	X		X	X				X						
54	M	19,1	C3		X		X		X vol+			X		X opr		X		
55	F	9		X	X		X	X				X	X			inegmi	X	mégagrde citerne
56	M	18	C2		X		X			X vol+		X	XX					
57	F	19	C3		X		X		X vol+			X	X	X		inégmiop	X	
58	F	3,5		X	X		X	X										
59	F	2		X	X		X	X										angiome plan joue
60	M	12		X	X			X				X						
61	M	14,1 X			X		X		X 1opr								X	
62	M	16,5	C1		X		X		X			X	XX	X dos			X	
63	M	5	C1		X		X	X				X	X		X	Xstab	X	
64	F	12,4 X			X		X			X		X					X	
65	M	14,7		X	X		X		X			X		Xthorax			X	
66	F	6,7 X			X		X	X									X	
67	F	9,1 X			X		X		X					X axill G			X	
68	F	7,1		X	X		X	X									X	
69	F	5,3		X	X		X	X				X(1)						
70	M	17,9		X	X		X			X		X		Xfrontal	X	pseu		sténose acq Sylvius
71	F	11 X			X		X	X				X					X	
72	M	15 X			X		X			X							X	
73	F	9		X	X		X		X			X					X	
74	F	4,5 X			X			X									X	

GLOSSAIRE

6 Sigma : Mesurer, analyser, améliorer et contrôler l'aptitude des processus afin de diminuer leur variabilité. Le «sigma» est une mesure de la variabilité des processus et le «six» indique jusqu'à quel niveau on doit réduire cette variabilité.

Analyse de la variance (Anova) : Technique statistique qui examine les différences entre au moins deux moyennes des éléments de classes afin de déterminer la pertinence de l'hypothèse que ces différences sont significatives.

Analyse discriminante : Technique statistique qui permet de prédire l'appartenance de sujets à une classe à partir de données sur une ou plusieurs variables continues.

Analyse exploratoire : Analyse de données destinée à découvrir des relations inconnues initialement entre les variables ou les données elles-mêmes.

Analyse factorielle : Technique statistique consistant à identifier des « facteurs » ou combinaisons d'un ensemble de variables mesurées permettant d'expliquer les relations mutuelles entre ces variables.

Apprentissage supervisé : Techniques d'apprentissage par lesquelles on cherche à approximer une variable donnée. Par exemple les techniques de régression ou de classification supervisées.

Apprentissage non supervisé : Techniques d'apprentissage pour lesquelles on recherche des regroupements de données sans faire appel à une variable de sortie particulière. Les techniques de « clustering » ou de classification non supervisée en font partie.

Arbre de décision : Méthode de représentation sous forme d'arbre d'un ensemble de tests hiérarchiques permettant de donner une valeur ou une classe comme résultat.

Associations : Techniques qui déterminent des règles conjonctive de la forme X^Y->A^B Les algorithmes d'association trouvent en général les associations qui satisfont un critère de support minimum et un seuil de confiance donné.

Base de règles, base de connaissances : Base de données contenant des connaissances tacites sous forme de faits codifiés de façon formelle et de règles de décision (si, alors, autrement (« if », « then », « else »)

Business Intelligence : Décrit un ensemble de concepts et de méthodes afin d'améliorer la prise de décision dans les entreprises en utilisant des systèmes d'aide basés en particulier sur les méthodes OLAP.

Carte de contrôle : Permet de contrôler et suivre l'évolution d'un processus en représentant graphiquement sa variabilité. Les cartes de contrôle permettent d'anticiper les dérives des processus tout en s'assurant que la production reste à l'intérieur de limites préétablies.

CIM : La Production Intégrée par Ordinateur (Computer Integrated Manufacturing) désigne l'ensemble des outils informatiques contribuant à l'automation et la supervision des moyens de production industriels.

Classification : Problème consistant à prédire une variable de sortie discrète à partir d'un modèle utilisant un certain nombre de variables d'entrée.

Clustering : Les algorithmes de clustering (ou classification non supervisée) recherchent et regroupent les données similaires d'une base de données. Par exemple, des clients peuvent être regroupés selon leurs comportements d'achat.

Corrélation : Relation mathématique entre deux variables qui s'influencent mutuellement. Deux variables qui varient ensemble dans la même direction ont une corrélation positive. Si elles varient dans des directions opposées, la corrélation est négative. Une corrélation nulle dénote l'absence de relation entre deux variables.

CRM : Outil de gestion de la relation client. Outil informatique développé pour permettre à une entreprise de fidéliser ses clients et d'accroître sa part du marché, en intégrant la gestion des données relatives aux besoins et aux attentes du client.

Data Mining : Activité d'extraction d'information dont le but est de découvrir des faits ou des connaissances dans les données. Le data mining utilise l'analyse statistique, des techniques d'extraction de connaissances, des techniques de modélisation pour trouver des relations entre les données et en déduire des modèles permettant d'effectuer des prédictions.

Data Warehouse : Synonyme de "entrepôt de données". Données provenant de différentes sources (systèmes de production, applications diverses, sources externes) et stockées dans une base de données unique pour fournir à l'utilisateur une vue intégrée et transversale des informations de l'entreprise.

DMAIC : Méthodologie de résolution de problèmes basée sur les 5 étapes suivantes : (1) Define : définir les buts du projet et les délivrables pour le client (2) Measure : mesurer le procédé pour déterminer la performance actuelle (3) Analyze : analyser pour déterminer les causes des problèmes constatés (4) Improve : améliorer le procédé pour éliminer les problèmes (5) Control : contrôler la performance future du système

Données manquantes : Certaines valeurs des bases de données peuvent être manquantes parce qu'elles n'ont pas été mesurées, renseignées, étaient inconnues ou ont été perdues. Certaines techniques ignorent les exemples pour lesquels une ou plusieurs valeurs sont ignorées, d'autres tentent de les reconstituer, certaines enfin tiennent compte de ces données en tant que données manquantes.

ERP : (Enterprise Resource Planning) progiciel de planification des ressources. Logiciel qui permet de gérer l'ensemble des processus d'une entreprise, en intégrant l'ensemble des fonctions de cette dernière comme la gestion des ressources humaines, la gestion comptable et financière, l'aide à la décision, la vente, la distribution, l'approvisionnement.

Knowledge Discovery : Processus consistant à identifier des informations et des relations valides, nouvelles, potentiellement utiles, et compréhensibles dans les données (définition donnée en 1976 par Fayyad, Piatetsky-Shapiro, et Smyth dans « Advances in Knowledge Discovery and data mining »).

K plus proches voisins : méthode de classification classifiant un point en calculant les distances de ce point aux différents points de l'ensemble des points d'apprentissage. Il affecte alors le point à classer à la classe la plus courante parmi ses k plus proches voisins (où k est un nombre entier).

Modèle : Représentation d'un phénomène particulier et des relations entre les variables pertinentes. Un modèle explicatif permet de comprendre le processus considéré ou son comportement. Un modèle prédictif permet de prédire une valeur inconnue de la variable de sortie en fonction des variables d'entrée.

Modèle Paramétrique / non-paramétrique : On parle de modélisation paramétrique lorsque le modèle est caractérisable par un vecteur de dimension finie et pas trop grande en pratique, et de modélisation non-paramétrique dans le cas contraire.

MSP : Maîtrise statistique des procédés : approche de maîtrise de la qualité en production basée en particulier sur l'utilisation des cartes de contrôle.

OLAP : "On-Line Analytical Processing" : outils donnant la possibilité à l'utilisateur de procéder à des analyses multi-dimensionnelles sur ses données, en effectuant des « coupes » sur ses données à l'aide de requêtes sur plusieurs attributs. Les moteurs OLAP utilisent en général des structures de données intermédiaires pour stocker des résultats pré-calculés, permettant ainsi un accès rapide aux résultats.

Prédicat : Une règle est composée de prédicats et de conclusions. Par exemple, la règle « Si T>10 et V=Tr Alors P=Vrai » comporte deux prédicats (« T>10 » et « V=Tr ») et une conclusion (« P=Vrai »).

Régression logistique : Technique généralisant la régression linéaire, utilisée pour prédire des variables discrètes.

Réseau Bayésien : Modèle probabiliste basé sur un graphe décrivant des probabilités conditionnelles entre des évènements.

Réseau neuronal : technique de modélisation non linéaire basée sur une analogie avec le fonctionnement des neurones du système nerveux. Un réseau neuronal peut être utilisé pour prédire des variables de sortie à partir de variables d'entrée.

Test d'hypothèse : Processus qui consiste à vérifier si les résultats obtenus auprès d'un échantillon peuvent être généralisés à la population dont provient cet échantillon. Syn.: Vérification d'hypothèse.

BIBLIOGRAPHIE

U.M. Fayyad and G. Piatesky-Shapiro and P. Smith. (1996) From Data mining to Knowledge Discovery : *An Overview. Advances in Knowledge Discovery and datamining, MIT Press*, 1-34

Blum and Mitchell (1998). Combining Labeled and Unlabeled Data with co-training : *COLT : Proceedins of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*

Quinlan J.R. C5.0 Datamining Tool. www.rulequest.com, 1997

Dieterich, T.G. (1998). Approximate statistical test for comparing supervised classification learning algorithms : *Neural computation, 10, 1895-1923*

Le Beux P, Jacquelinet C. Intégration des hôpitaux universitaires aux réseaux de la recherche : une nécessité. In: *Informatique et Santé*. Springer Verlag France, 1993, 6:45-57.

Berners-Lee TJ, Cailliau R, Groff JF, Pollermann B, CERN. World-Wide Web:The Information Universe. *Electronic Networking : Research, Applications and Policy*. 1992; 2(1):52-8.

Sowa JF. Conceptual structures. *Information processing in Mind and Machine*. Reading, Massachusetts: Addison Wesley, 1984.

Warner HR, Toronto AF, Veasy LG. Experiences with Bayes theorem for computer diagnosis of congenital heart diseases. *Am. N Y Academ. Sci.* 1964; 115:538-67.

De Dombal FT, Leaper D, Staniland JR, McCann AP, Horrocks JC. Computer aided diagnosis of acute abdominal pain. *Br. Med. J.* 1972; 2:9-13.

Shortiffe EH, Scott AC, Bischoff MB, Campell AB, Van Melle W, Jacobs CD. An expert system for oncology protocol management. In: *Rule-Based Expert Systems - The MYCIN Experiments of the Stanford Heuristic Programming Project*. Buchanan BG and Shorliffe EH (eds). Reading, Massachusetts:Addison-Wesley, 1984:653-65.

Kulikovski CA, Weiss SM. Representation of expert knowledge for consultation: the CASNET and Expert projects. In: *Artificial intelligence in Medicine*. Szolovits P (ed). Westview:Press, 1982.

Miller RA, Pople HE, Myers JD. Internist-I. An experimental computer-based diagnostic consultant for general internal medicine. *N Engl Med*. 1982; 307:468-77

Miller R, Masarie FE, Myers JD. Quick Medical reference (QMR) for diagnostic assistance. MD Comput. 1986; 3:34-48.

Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. Dxplain : Anevolving diagnostic decision-support system" *Jama*. 1987; 3:258.

Fieschi M. In: *Intelligence Artificielle en médecine : des systèmes experts*. Paris:Masson. 1984.

Le Beux P, Fontaine D. Un système d'acquisition de connaissances pour systèmes experts". *Rev Tech Sciences Inform.* 1986; 5(1):7-20.

Lincoln M, Turner C, Haug P, Warner H, Williamson J, Bouhaddou O. Iliad training enhance medical students diagnostic skills *J Med Syst.* 1991.

Morice V, Seroussi B, Charlet J, Boisvieux JF. Représentation et gestion des connaissances pour une aide informatisée à la prescription et à la surveillance. *Informatique et Santé.* 1992; 5:33-45.

AIM-CEN *European Workshop on Medical Record Vol II.* Brussells:March April 1993.

Heathfield HA, Wyall J. Philosophies for the design and development of clinical decision-support systems. *Meth Inform Med.* 1993; 32:1-8.

Lindberg DAB. High performance computing and communications : the medical connection. In:*Proc MIE93.* Freund Pub. Home Ltd. 1993; 19-23.

Jaulent MC, Jean FC, Sauquet D, Degoulet P. The interface manager of the HELIOS medical software engineering environment. In : Lun KC, Degoulet P, Piemme TE, Rienhoff O (eds). *MEDINFO 92.* North-Holland. 1992 ; 1362-67.

Baud RH, Rassinoux AM, Scherrer JR. Natural language processing and medical records. In : Lun KC, Degoulet P, Piemme TE, Rienhoff O (eds). *MEDINFO 92.* North-Holland. 1992 ; 1362-67.

Rassinoux AM, Baud RH, Scherrer JR. Conceptual graphs model extension for knowledge representation of medical texts. In : Lun KC, Degoulet P, Piemme TE, Rienhoff O (eds). *MEDINFO 92.* North-Holland. 1992 ; 1362-67.

Pearl J. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence.* 1986; 29:241-88

Musen MA. Dimensions of knowledge sharing and reuse. *Comp Biomed Res.* 1992; 25(5):435-467.

Hripcsak G, Clayton PD, Pryor TA, Haug P, Wigertz OB, Van der Lei J. The Arden syntax for Medical Logic Modules. In: Miller RA (ed). *Proc 14th SCAMC.* Washington: IEEE Comput Soc Press, 1990; 200-4.

Stefanelli M. European research efforts in medical knowledge-based systems. *Artificial Intelligence Med.* 1993; 5:107-24.

Musen MA, Fagan LM, Combs DM, Shortliffe EH. Use of a domain model to drive an interactive knowledge-editing tool. *Int J Man-Machine Stud.* 1987; 26:105-21.

Patil RS, Szolovits P, Schwartz WB. Modeling Knowledge of the Patient in Acid-Base and Electrolyte Disorders. In: *Artificial Intelligence in Medicine.* Szolovits P. ed. Wetview: Press. 1982; 191-226.

Ermine JL, Cauhapé D. Raisonnement temporel dans les systèmes experts. *Rev Intelligence Artificielle*. 1990; 4(1):99-136.

Cousins SB, Chen W, Frisse ME. A tutorial introduction to Stochastic Simulation Algorithms for Belief Networks. *Artificial Intelligence Med.* 1993; 5:315-40.

Pearl J. In: *Probabilistic Reasoning in Intelligent Systems*. San Marteo: Morgan Kaufmann Publishers. 1988.

Andreassen S, Jensen FV, Olesen KG. Medical expert systems based on causal probabilistic networks. *Int J Biomed Comp*. 1991; 28:1-30.

Heckerman DE, Nathwani BN.Toward Normative Expert Systems : Part II. Probability-Based Representations for Efficient Knowledge. Acquisition and Inference. *Meth inform Med.* 1992; 31(2):106-16.

Lauritzen SL, Spiegelhalter DJ Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J Royal Stat Soc B.* 1988; 50:157-224.

Lauritzen SL. *The EM algorithm for graphical association models with missing data.*. Aalborg: Aalborg university. Technical report R 91-05. 1991.

Neapolitan RE. Computing the confidence in a medical decision obtained from an influence diagram *Artificial Intelligence Med* . 1993; 5:341-63.

Xiang Y, Pant B, Eisen A, Beddoes MP, Poole D. Multiply sectioned Bayesian networks for neuromuscular diagnosis. *Artificial Intelligence Med* . 1993; 5:293-314.

Levitt TS, Hedgcock MW, Dye JW, Johnston SE, Shadle VM, Vosky D. Bayesian inference for model-based segmentation of computed radiographs of the hand. *Artificial Intelligence Med* . 1993; 5:365-87.

Wagner HM. *Principles of operation research*. (second edition). Prentice Hall 1975. [58] Colmerauer A. Opening the Prolog III Universe. *Byte Magazine Special Issue on Logic Programming.* Aug 1987.

Weil GE. L'Hopital au service du malade : transfert des concepts, des méthodes, des outils de la gestion de production. *Thèse de Génie Biologique et Médical.* Université de Grenoble. Mars 1990.

Widmer G, Horn W, Nagele B. Automatic knowledge base refinement : learning from examples and deep Knowledge in rheumatology. *Artificial Intelligence Med.* 1993; 5:225-44.

Lanzola G, Stefanelli M. Inferential Knowledge acquisition. *Artificial Intelligence Med.* 1993; 5:253-68.

Berman L, Cullen M, Miller PL. Automated Integration of External Databases:A knowledge based Approach to Enhancing Rule Based Expert system*Comp Biomed Res.* 1993; 26(3):230-41.

Giuse DA, Giuse NB, Miller RA.Consistency enforcement in medical knowledge base construction. *Artificial Intelligence in Med.* 1993; 5:245-52.

McCulloch W.S, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of mahematical Biophysics.* 1943; 5:115-133.

Hebb DO. The organization of behaviour. *J. Wiley ans sons* . 1949.

Reggia JA. Multiple disorder diagnostis with adaptive competitive neural networks.*Artificial Intelligence Med.* 1993; 5(6): 469.

Masic N, Pfurtscheller G. Neural networks based classification of single-trial EEG data. *Artificial Intelligence Med.* 1993; 5(6): 503.

Barreto JM, De Azevedo FM. Connectionist expert systems as medical decision aid. *Artificial Intelligence Med.* 1993; 5(6): 515

De Moor GJE. Standardization in Medical informatics. *Yearbook Med Inform* 1993:61-6.

Kornblum C. L'informatique médicale et hospitalière. Enjeux médicaux et industriels *Rapport Ministère de l'Industrie* : juillet 1992

Cawthon RM, Weiss M, Xu G, et al. A major segment of the neurofibromatosis type 1 gene: cDNA sequence, genomic structure, and point mutations. *Cell.* 1990;62:193-201.

Viskochil D, Buchberg AM, Xu G, et al. Deletions and a translocation interrupt a cloned gene at the neurofibromatosis type 1 locus. *Cell.* 1990;62:187-192.

Riccardi VM. Type 1 neurofibromatosis and the pediatric patient. *Curr Probl Pediatr* 1992 Feb;22(2):66-106; discussion 107.

National Institutes of Health Consensus Development Conference. Neurofibromatosis: Conference Statement. *Arch Neurol.* 1988;45:575-578.

Obringer AC, Meadows AT, Zackai EH. The diagnosis of neurofibromatosis 1 in the child under the age of 6 years. *Am J Dis Child.* 1989;143:717-719.

Listernick R, Charrow J. Neurofibromatosis type 1 in childhood. *J Pediatr* 1990 Jun;116(6):845-53.

Crowe FW, Shull WJ, Neel JV. A clinical, pathological and genetic study of Multiple Neurofibromatosis. 1956, Charles C. Thomas, Springfield, Illinois.

Riccardi VM, Mulvihill jj, eds. Neurofibromatosis (Von Recklinghausen disease). Adv Neurol. 1981 ;29 :1-282.

Riccardi VM, Eichner JE. *Neurofibromatosis: Phenotype, Natural History and Pathogenesis.* 2nd ed. Baltimore, Md: Johns Hopkins University Press; 1992.

Friedman JM, Birch P, Greene C. National Neurofibromatosis Foundation International Database. *Am J Med Genet* 1993 Jan 1;45(1):88-91.

Crowe FW, Schull WJ. Diagnostic importance of the café au lait spot in neurofibromatosis. Arch ; Intern. Med.1953 ;91 :758-66.

Burwell RG, James NJ, Johnston DI. Cafe-au-lait spots in schoolchildren. *Arch Dis Child* 1982 Aug;57(8):631-2.

Gutmann DH, Aylsworth A, Carey JC, Korf B, Marks J et col. The diagnostic evaluation and multidisciplinary management of neurofibromatosis 1 and neurofibromatosis 2. *JAMA* 1997 Jul 2;278(1):51-7.

Korf BR. Diagnostic outcome in children with multiple cafe-au-lait spots. *Pediatrics.* 1992;90:924-927.

Huson S, Jones D, Beck L. Ophthalmic manifestations of neurofibromatosis. *Br J Ophthalmol* 1987 Mar;71(3):235-8.

Huson SM, Harper PS, Compston DAS. Von Recklinghausen neurofibromatosis: a clinical and population study in South East Wales. *Brain.* 1988;111:1355-1381.

North K. Neurofibromatosis type 1: review of the first 200 patients in an Australian clinic. *J Child Neurol.* 1993;8:395-402.

Listernick R, Louis DN, Packer RJ, Gutmann DH. Optic pathway glioma in children with neurofibromatosis 1: consensus statement from the Optic Pathway Glioma Task Force. *Ann Neurol.* 1997;41:143-149.

Listernick R, Darling C, Greenwald M, Strauss L, Charrow J. Optic pathway tumors in children: the effect of neurofibromatosis 1 on clinical manifestations and natural history. *J Pediatr.* 1995;127:718-722.

Lewis RA, Gerson LP, Axelson KA, Riccardi VM, Whitford RP. von Recklinghausen neurofibromatosis. II. Incidence of optic gliomata. Ophthalmology 1984 Aug;91(8):929-35.

Es SV, North KN, McHugh K, Silva MD. MRI findings in children with neurofibromatosis type 1: a prospective study. Pediatr Radiol 1996 Jul;26(7):478-87.

Crossett LS, Beaty JH, Betz RR, Warner W, Clancy M, Steel HH. Congenital pseudarthrosis of the tibia. Long-term follow-up study. *Clin Orthop* 1989 Aug;(245):16-8.

Stevenson DA, Birch PH, Friedman JM, Viskochil DH, Balestrazzi P, Boni S et al. Descriptive analysis of tibial pseudarthrosis in patients with neurofibromatosis 1. *Am J Med Genet* 1999 Jun 11;84(5):413-9.

Chapman CA, Waber DP, Bassett N, Urion DK, Korf BR. Neurobehavioral profiles of children with neurofibromatosis 1 referred for learning disabilities are sex-specific. *Am J Med Genet.* 1996;67:127-132.

North KN, Riccardi V, Samango-Sprouse C, Ferner R, Moore B, Legius E et al. Cognitive function and academic performance in neurofibromatosis. 1: consensus statement from the NF1 Cognitive Disorders Task Force. *Neurology* 1997 Apr;48(4):1121-7.

TABLE DES MATIERES

SOMMAIRE p.2

NOM : COUPVENT DES GRAVIERS **PRENOM** : ANTOINE

TITRE DE LA THESE :

UTILISATION D'UNE TECHNIQUE D'APPRENTISSAGE NON PARAMETRIQUE SUR UNE BASE DE DONNEES DE 87 ENFANTS ATTEINTS DE NEUROFIBROMATOSE DE TYPE 1

RESUME

Parmi les nouvelles méthodes d'exploration de bases de données médicales, les algorithmes d'apprentissage non paramétrique présentent l'avantage de s'extraire de la contrainte d'un modèle déjà établi par l'expert.

Nous avons appliqué ce type d'algorithme sur une base de données de 87 enfants atteints de neurofibromatose de type 1 et avons mis en évidence les points suivants :

- Tout enfant dans la base de données d'âge inférieur à 12 ans et ne présentant pas d'éphélides n'aura pas de difficultés d'apprentissage ;
- Il n'existe aucune règle pertinente dans cette base de données, permettant de corréler la présence ou non de pseudarthrose avec les autres symptômes de la maladie.
- Il n'existe, dans cette base de données, aucune règle prédictive du sexe agrégeant plus de 6 individus et restant pure à 100%.

MOTS-CLES

BASE DE DONNEES MEDICALE
NEUROFIBROMATOSE
ALGORITHME D'APPRENTISSAGE NON PARAMETRIQUE
EXTRACTION DE REGLES