II QU'EST-CE QUE LA ROBUSTESSE ?

Dans le premier chapitre, nous avons fait l'historique de notre démarche fondée sur l'utilisation des différentes solutions adaptées à la construction de systèmes experts en pathologie végétale.

Nous voici maintenant devant un nouveau problème de classification et de détermination dans le domaine de la systématique. Au départ, nous avions à notre disposition un logiciel d'apprentissage automatique de règles de décision à partir d'exemples : KATE [Manago & Conruyt, 1989]. Si nous savions comment opérer avec les exemples (par induction), l'utilisation de cet outil supposait préalablement réglées deux questions importantes :

1) Quelles descriptions traiter ?
2) Comment les acquérir ?

Une troisième difficulté a été identifiée lors de résultats expérimentaux dans une application de détermination d'objets militaires [Manago, 1991]. En présence d'observations incomplètes (dues au camouflage par exemple), le système expert engendré par KATE pouvait fournir un diagnostic incertain et ne pas lever l'ambiguïté entre un char et de l'artillerie légère !

Les deux premières questions sont de nature qualitative : la qualité des exemples à apprendre est une *caractéristique* importante avant leur traitement ; elle dépend du bon déroulement de la *procédure* de description elle-même. Nous montrerons ce premier aspect de la **robustesse de la description** dans ce chapitre.

Ensuite, nous relierons la troisième question à la **robustesse de la consultation** face aux valeurs manquantes ou réponses «inconnu». Nous verrons dans cette thèse comment nous sommes parvenus à répondre à ces différentes questions sur la robustesse. Mais, auparavant, nous allons étudier ses différents aspects théoriques et pratiques ainsi que ses diverses interprétations dans la communauté scientifique et parmi les utilisateurs.

2.1 Aspects théoriques

2.1.1 La robustesse statistique

L'étude statistique d'une base d'exemples vise à produire un résumé d'un fichier de centaines exemples décrits par des dizaines de variables. Ce résumé prend la forme d'un arbre dont chaque nœud correspond à une partie des exemples ayant les mêmes valeurs pour certaines variables. De même qu'un histogramme est une image qui résume un fichier uni-colonne, un arbre est avant tout une image résumée d'un fichier multi-colonnes correspondant à des variables n'ayant qu'un petit nombre de valeurs [Crémilleux, 1991]. Le processus par lequel on synthétise les exemples est appelé induction.

Pour résumer l'information, les systèmes d'apprentissage inductif recherchent des régularités dans les données d'observation initiales en utilisant des critères numériques issus des statistiques (χ^2, critère de Gini, entropie de Shannon, etc.), ce qui permet de prendre des choix décisifs pour partitionner les exemples. La séparation est censée avoir une signification statistique, c'est-à-dire qu'elle ne découle pas simplement du hasard [Gascuel & Carraux, 1992]. Le principe de construction des arbres de décision est expliqué au chapitre 7.

L'objectif des statisticiens est d'utiliser ces arbres comme un moyen efficace de prédire le classement de nouvelles observations avec un taux minimal d'erreurs. C'est le pouvoir prédictif de l'arbre qui détermine sa robustesse statistique dans ce contexte [Breiman *et al.*, 1984]. Une recherche de Mingers sur des données empiriques [Mingers, 1989] aboutit à la conclusion que ce n'est pas tant le choix de la mesure qui importe mais plutôt celui de l'élagage de l'arbre final. Ainsi, le programme CART extrait le meilleur sous-arbre en utilisant soit un critère d'élagage pour les grosses bases d'exemples, soit une validation croisée lorsqu'il y a peu d'exemples [Gomes, 1992].

La robustesse statistique suppose néanmoins certaines hypothèses probabilistes posées *a priori* de manière à pouvoir estimer la reproductibilité des résultats de classement des nouvelles observations :

◊ la représentativité de la base d'exemples nécessite de considérer la fréquence d'apparition des exemples dans la population, les cas rares n'ayant pas le même poids statistique que les cas "typiques"[1],

◊ l'échantillonnage se fait de manière aléatoire en suivant un modèle de distribution de la population étudiée.

[1] Pour une explication des différents sens du terme "typique", on peut se référer à [Lebbe, 1991].

2.1.2 Le formalisme mathématique de description

La conception d'outils informatiques adaptés aux problèmes des biologistes nécessite à la fois l'utilisation de techniques en statistiques, en analyse de données, en intelligence artificielle, en ergonomie et en psychologie cognitive. Le sujet se situant à la frontière de ces différents domaines, il est important de présenter formellement les problèmes tels qu'ils se posent aux biologistes de manière à pouvoir faire comprendre leur nature aux différentes communautés amenées à les résoudre. Le langage mathématique est ainsi le dénominateur commun permettant une meilleure communication entre les personnes concernées et se trouve par conséquence être un facteur important de la robustesse des solutions apportées. C'est pourquoi le chapitre 5 expose le formalisme mathématique de description des sujets étudiés au MNHN, ce même formalisme étant ensuite exploité dans le chapitre 7 pour la description des algorithmes de traitement des exemples.

2.1.3 Combiner du numérique et du symbolique

L'approche numérique qui est utilisée dans le traitement permet de discriminer efficacement un grand nombre d'exemples tout en tenant compte des petites variations dans les descriptions. Elle permet aussi de détecter un type de bruit particulier ou deux exemples portent la même description tout en n'appartenant pas à la même classe : on a alors à faire à une ambiguïté totale, ce qui laisse supposer à l'utilisateur que les mêmes causes ne produisent pas les mêmes effets. Or, la mise en évidence d'un tel "clash" (Crémilleux, 1991) peut faire réagir l'expert : il peut s'apercevoir qu'il a oublié de décrire un caractère discriminant entre les deux exemples (désambiguation).

L'approche symbolique permet de représenter des connaissances complexes en indiquant les dépendances entre objets, attributs et valeurs ainsi que des règles de cohérence pour chaque description. Elle donne aussi la possibilité d'introduire des connaissances complémentaires aux exemples pour traiter certains bruits (voir plus loin) [Manago, 1988]. En les explicitant, l'apprentissage symbolique fournit des explications justifiées par la présence de connaissances non fortuites [Kodratoff, 1991].

L'intégration des deux approches améliore la robustesse globale du système.

2.2 Aspects pratiques

Nous avons déjà donné une définition de la **robustesse** que nous qualifions d'**empirique** car basée sur les pratiques des utilisateurs : c'est l'ensemble des facteurs qualitatifs qui améliore l'acquisition des connaissances sur le domaine ou encore permet d'éliminer certaines faiblesses liées à l'utilisation des outils.

2.2.1 Les facteurs qualitatifs

2.2.1.1 Fiabilité

Dans le cadre de la validation des systèmes experts en pathologie des plantes à l'INRA, nous avons évalué la fiabilité des **résultats d'identification** lorsque les programmes sont mis dans les conditions normales d'utilisation, c'est-à-dire entre les mains des techniciens et des agriculteurs qui n'ont pas la même manière d'observer que l'expert.

Dans le cadre de l'apprentissage, nous avons constaté que la fiabilité des résultats dépendait surtout de la qualité des **données** en entrée (§ 1.3.4). Nous nous sommes alors attaqués en priorité à la robustesse de l'acquisition des connaissances, c'est-à-dire non pas à celle des règles élaborées par méthode d'élicitation comme pour les systèmes experts de première génération, mais à celle des données initiales sur lesquelles va s'opérer l'induction : on suppose que le traitement qui suivra, s'il est bien justifié, donnera des règles et des résultats fiables par rapport aux données robustes.

Les données en entrée sont de deux sortes : les premières sont des **connaissances observables** et générales sur le domaine, représentées dans le modèle descriptif. Les secondes sont des **connaissances observées** spécifiques, correspondant aux exemples d'apprentissage. Cette distinction au niveau des descriptions est fondamentale pour évaluer leur fiabilité.

2.2.1.2 Compréhension

Pour obtenir des données robustes, il est nécessaire de bien comprendre le domaine. Ceci est d'abord vrai au niveau de la compréhension entre l'expert et le cogniticien ; le fait que ce dernier ait une compétence ou une sensibilité sur le domaine facilite grandement le dialogue. Mais surtout, comme les utilisateurs qui identifient des échantillons ont des niveaux de connaissance très variés sur le sujet, la phase de modélisation pour acquérir l'observable est un travail d'équipe essentiel entre l'expert et le cogniticien. Le but est de réfléchir sur les aspects terminologiques afin de trouver une structure de description des composants du

domaine qui soit cohérente, bien comprise et bien interprétée par les utilisateurs ciblés. Il s'agit d'une chasse aux ambiguïtés de toute nature.

L'adaptation au niveau de compréhension de l'utilisateur est un facteur important de la robustesse. Par exemple, pour que des douaniers utilisent efficacement un système expert d'identification des espèces menacées d'extinction, il leur faut un guide d'observation et un **vocabulaire** adapté pour se familiariser avec les critères souvent pointus de discrimination entre deux espèces (l'une protégée par la convention de Washington et l'autre non). Ces personnes "naïves" par rapport à l'observation utiliseront d'autant mieux le questionnaire de saisie des descriptions que celui-ci est bien structuré, des dessins explicatifs illustrant le vocabulaire spécialisé.

2.2.1.3 Précision

La précision intervient dans le degré de finesse du processus de classification et/ou d'identification. C'est pourquoi il convient de fixer des limites au niveau des détails de description à fournir au niveau de *l'observable*. Les descriptions détaillées dépendent des techniques d'observation possibles au moment de l'identification. Par exemple, pour reconnaître des espèces d'Hydraires, il peut être avantageux d'utiliser les possibilités d'observation du microscope à balayage électronique si les utilisateurs ont accès à ce type de matériel.. Cela donne la possibilité d'introduire des caractères internes de différentiation des nématocystes (capsules urticantes) dans le modèle descriptif. Mais on peut aussi se contenter des formes extérieures de ces mêmes composants qui ne nécessitent qu'une observation au microscope optique (au plus fort grossissement toutefois). Le choix est un compromis opérationnel qui dépend des objectifs de la description et des moyens disponibles pour l'observation.

La précision est aussi un facteur que l'on peut rapprocher de la **justesse** des descriptions *observées*. Ces dernières doivent représenter fidèlement la réalité des échantillons au moment de leur saisie dans le questionnaire.

2.2.1.4 Exhaustivité

Une fois fixés les objectifs, l'exhaustivité des caractères mis en jeu dans le modèle descriptif est alors très importante. Nous pouvons alors cerner le problème observable, nous assurer de sa complétude par rapport au domaine qui a été bien délimité, et ainsi répertorier les valeurs admissibles dans le questionnaire. L'exhaustivité au niveau de l'observable implique de fournir à l'expert une certaine souplesse d'expression, avec un langage de représentation des connaissances suffisamment puissant : logique multi-valuée, avec variables (ordre 1), taxonomie de valeurs, démons entre objets du modèle, etc.. Pour les utilisateurs, le langage est néanmoins rendu transparent au niveau syntaxique par

une interface de saisie conviviale. En outre, il est bon de favoriser l'expression sémantique des caractères, leur interdépendance, le choix judicieux des valeurs possibles par rapport à la signification de l'attribut (**monosémie** des caractères).

L'exhaustivité doit se concrétiser aussi au niveau des descriptions observées qui devraient être **complètes** par rapport à l'échantillon disponible. Par exemple, il est bon d'indiquer à l'utilisateur d'éviter les idées préconçues sur le diagnostic de l'échantillon : il s'agit d'un biais qui le polarise sur la description des symptômes correspondants. La règle serait d'éviter que l'utilisateur décrive ce qu'il cherche plutôt que ce qu'il peut voir sur la plante !

2.2.1.5 Cohérence

Notre objectif est d'assurer une certaine cohérence du modèle descriptif au niveau de la définition du statut des caractères (“objet-attribut-valeur”), ainsi que dans celle des relations entre les objets observables (objets de type composant, point-de-vue, spécialisant). Ce facteur oblige l'expert à plus de rigueur et de rationalité dans sa manière de structurer son modèle descriptif (par exemple en appliquant la règle de définir les objets du plus général au plus précis).

Une fois ce travail accompli, une autre cohérence intervient en phase d'acquisition des exemples à apprendre : c'est celle de l'ajustement de l'observé par rapport à l'observable. Elle permet d'**éviter les oublis éventuels** non perçus lors du remplissage du questionnaire. En effet, lorsqu'il s'agit de passer de l'observable à l'observé (le modèle descriptif servant de moule à la constitution d'un questionnaire “guide d'observation”), tous les caractères (objets et attributs) seront passés en revue lors d'une consultation pour que l'utilisateur puisse affirmer soit leur présence (ou absence), soit le fait que l'on ne peut pas les renseigner (réponse «inconnu»). Au départ de la description, chaque caractère est sans statut (présent, absent ou inconnu). La vérification des oublis doit intervenir à la fin lorsque l'utilisateur indique qu'il a fini sa description : elle est appliquée pour assurer la cohérence de l'utilisateur vis-à-vis de ses réponses (différence entre l'oubli et l'inconnu).

2.2.1.6 Redondance

Ensuite, nous mentionnerons le rôle de la redondance dans la représentation de la diversité de l'observé. En effet, pour nos classifications biologiques, l'exception a autant d'importance que le cas général pour découvrir et caractériser le *continuum* entre les Espèces. Le cas particulier n'est pas un biais à éviter mais plutôt une richesse à représenter dans les descriptions. Pour une classe donnée, nous souhaiterons acquérir sa couverture la plus large possible en nombre d'exemples. Cela correspond à la vision extensive de la définition d'une classe ou encore définition d'un concept du point de vue des exemples [Smith &

Medin, 1981]. L'objectif est donc de multiplier le nombre de descriptions d'une même classe même si elles se ressemblent fortement. Cette manière de procéder n'est pas superflue du fait de la variabilité naturelle observée au niveau des spécimens.

2.2.1.7 Mise à jour

Comme il n'est pas possible pratiquement de tout prévoir dès le départ dans le modèle descriptif, la mise à jour des connaissances est un facteur de robustesse à prendre en compte obligatoirement. Par exemple, des nouvelles maladies apparaissent tous les ans en pathologie végétale ou encore une maladie déjà répertoriée montre des symptômes différents une certaine année. Le but est de savoir maintenir la base d'exemples en fonction des modifications apportées dans le modèle. Ce facteur est à relier au critère plus global d'incrémentalité temporelle (voir plus loin). Il donne tout son sens à la robustesse empirique dont la nature évolutive est fondée sur la découverte et l'interactivité avec l'expert. De son coté, la robustesse statistique se concentre plus sur les conditions de la reproductibilité des résultats de classification. Ces deux aspects de la robustesse ne sont pas incompatibles.

Néanmoins, la représentativité des données n'est pas un critère applicable dans le domaine de la systématique : nous avons affaire à relativement peu de données par classe (en nombre d'individus) par rapport au nombre de variables possibles : chaque individu est complexe à décrire. Dans ce contexte, la manière de les décrire est sujette à de multiples révisions.

Les modifications à apporter font suite à la procédure de validation des connaissances apprises. Elle intervient aussi bien après le traitement des données qu'au moment de l'aller-retour entre la définition du modèle descriptif et la saisie des exemples dans le questionnaire. La robustesse empirique procède de manière cyclique à l'aide de la mise à jour et va dans la direction d'une plus grande précision des résultats. Ce principe est de plus en plus à l'ordre du jour des recherches en apprentissage et en raisonnement à partir de cas [Utgoff, 1989], [Aamodt, 1989].

2.2.1.8 Ergonomie

Citons encore l'ergonomie qui est tout ce qui facilite l'utilisation des outils (modèle descriptif, questionnaire, système expert) et rend la consultation plus agréable. Par exemple, la **convivialité** doit faciliter la communication entre la machine et l'utilisateur. L'**interactivité** est l'ensemble des fonctionnalités et des performances du système informatique qui permet la réalisation d'une tâche sans perturber le processus mental que l'utilisateur suit pour l'accomplir. C'est aussi

la capacité de l'utilisateur d'interrompre le raisonnement en cours et de garder le contrôle sur la machine [Bove & Rhodes, 1990].

Le but de la convivialité est d'obtenir un outil simple d'emploi. Cela peut être accompli grâce aux possibilités hypermédia (hypertexte, image, son, vidéo) du Macintosch ainsi que de programmes comme HyperCard avec son langage HyperTalk [Apple, 1988]. Ces outils permettent de représenter la connaissance de manière visuelle, chaque nœud ou objet du modèle étant symbolisé par une carte qui peut recevoir une image ou un dessin expliquant le concept et des boutons pour se déplacer vers d'autres objets. L'utilisateur n'a qu'à pointer sur l'objet désiré et cliquer pour y aller, ce qui est très naturel. Un intérêt est par exemple d'utiliser une palette de couleurs ou des dessins à la place du choix des valeurs (les mots) elles-mêmes par l'utilisateur. Il n'a plus qu'à cliquer sur la représentation visuelle au lieu d'interpréter le nom associé, ce qui peut provoquer des erreurs de description.

Néanmoins, l'ergonomie ne se résume pas seulement à employer des outils conviviaux (point de vue statique). Il faut savoir les utiliser à bon escient, organiser la connaissance pour satisfaire à l'objectif d'interactivité (point de vue dynamique). Par exemple, les nœuds sont reliés entre eux au sein d'une hiérarchie arborescente à explorer qui n'est pas un réseau sémantique multi-directionnel. La navigation est ainsi orientée par la volonté pédagogique de l'expert de guider l'observation selon un ordre bien établi (du général au particulier). Il pourra très facilement rajouter des explications ainsi que des messages d'aide à l'observation à l'aide de boutons (quoi faire, comment faire, mise en garde avant une action, alerte après, etc.) pour éduquer l'utilisateur.

Rada et Barlow (1989) ont gagé sur l'avenir de “l'expertexte” qui mixe les deux technologies des systèmes experts et de l'hypertexte. Nous y ajouterons simplement la technologie multimédia pour ses capacités ergonomiques et éducatives [Hooper, 1990].

Dans ce travail, nous juxtaposerons toutefois les deux approches sans les mélanger : nous utiliserons l'hypertexte avec HyperQuest dans le cadre de l'acquisition des descriptions en amont de la phase d'apprentissage. Le système expert engendré par KATE est un programme écrit en C et sa consultation aura lieu dans cet environnement. Les deux modules sont bien séparés. Ces deux applications communiquent leurs connaissances par l'intermédiaire de fichiers ASCII (modèle descriptif et exemples). Leur véritable intégration sera envisagée à la suite de cette thèse à l'aide des “Apple Events” qui autorisent la communication plus facilement entre les applications.

2.2.1.9 Tolérance aux bruits

Enfin, nous mentionnerons le facteur de robustesse qui nous parait le plus important : la tolérance aux bruits. Dans INSTIL, il y avait deux problèmes attachés au bruit : la **détection** et le **traitement**. Pour le premier aspect, les différentes sortes de bruit ont été identifiées et répertoriées au niveau des trois phases de l'acquisition des connaissances : collecte et observation, description, diagnostic. La classification de la figure 2.2 en donne un résumé (voir plus loin). Pour le second aspect, une bonne partie des bruits des différents maillons de la chaîne a pu être traitée avant la phase d'apprentissage afin d'obtenir des exemples de qualité. Les moyens à mettre en œuvre pour minimiser ces bruits «de terrain» sont décrits dans [Conruyt & Piaton, 1987].

Néanmoins, d'autres bruits plus «abstraits» sont par exemple la difficulté d'observation d'un caractère, son polymorphisme, son coût, la fiabilité du diagnostic, la tolérance d'une coupure autour d'un seuil d'une variable numérique, l'importance d'un caractère comme critère de classification. Ils nécessitent une représentation symbolique explicite dans les exemples pour leur traitement [Manago & Kodratoff, 1987]. Ce travail a été réalisé en introduisant des propriétés supplémentaires dans la définition des attributs [Manago, 1988], [Conruyt & Lesaffre, 1988] :

◊ Confiance
Ce paramètre définit simplement le coefficient de vraisemblance d'une information. Sa valeur sera “faible” si l'attribut est difficile à observer. Les attributs ayant un faible degré de confiance sont utilisés le plus tard possible durant la construction de l'arbre de décision.

◊ Recouvrement
Lorsque des valeurs se recouvrent, comme par exemple, [couleur tache (recouvrement (brun beige) (brun noir))], la sélection des exemples à un nœud de l'arbre de décision pour le test “couleur(tache)” tiendra compte de la polymorphie des couleurs : pour la valeur “noir”, on retiendra pour construire le sous-arbre tous les exemples dont la couleur de la tache est aussi “brun”.

◊ Coût
Ce paramètre indique le prix à payer (financier, temps d'attente, etc.) pour obtenir la réponse au test demandé. Par exemple, faire un test de laboratoire (isolement bactérien, viral) possède un coût élevé. On essayera donc d'abord les tests bon marché pour construire les règles de décision.

◊ Fiabilité
Il s'agit ici de la confiance que l'expert accorde au diagnostic d'un exemple. C'est une mesure de la qualité d'un exemple en terme de

diagnostic. Ce test a pu être utilisé par Main pour privilégier l'utilisation d'exemples fiables lors de la sélection du noyau.

◊ Tolérance
Les seuils numériques ont un caractère tranché qui ne convient pas toujours à la précision des mesures effectuées. On peut donc considérer qu'il existe une marge d'erreur possible autour de ce seuil qu'il est intéressant de spécifier. La tolérance est donc une mesure de recouvrement lorsque l'on compare des valeurs numériques. Elle peut être explicitée de manière relative ou absolue.

◊ Priorité
Ce dernier paramètre permet à l'expert d'influer sur la classification. Les caractères n'ont pas tous la même importance de son point de vue pour caractériser une classe ou un diagnostic. La prise en compte de la priorité de certains caractères peut se faire par exemple au niveau du calcul d'entropie pour classer les attributs ayant le même gain d'information au nœud courant.

Tous ces paramètres sont des connaissances symboliques supplémentaires qui tiennent compte des spécificités du domaine. Elles doivent être explicitées dans le modèle descriptif en fonction des besoins exprimés par l'expert.

2.2.1.10 Adaptation aux besoins exprimés

En ce qui concerne l'application des éponges marines au MNHN (§ 1.5.3), l'introduction de ces paramètres pour traiter ces différentes sortes de bruit n'est pas demandée. En particulier, il n'y a pas d'*a priori* sur la priorité d'un caractère pour construire une classification. Il faut dire que dans cette application, l'expert est à la fois professeur et descripteur, il n'y a pas une grande variabilité d'utilisateurs potentiels du système expert. La demande est plutôt celle d'adapter des outils d'aide à la classification au travail quotidien des biologistes systématiciens.

Ainsi, nous devrons nous adapter à la démarche naturelle de l'expert qui est la suivante :

1) observer et se familiariser,
2) représenter les observations => établir des **descriptions**,
3) bâtir des hypothèses à partir des descriptions (pré-classées ou non)
=> construire des règles de **classification**,
4) les éprouver par de nouveaux faits => conduire une **détermination**.

Nous chercherons donc à construire une méthode d'acquisition des connaissances qui s'appuie sur différents savoir-faire tels que les capacités

d'observation, de description et de raisonnement des systématiciens et qui tienne compte à la fois de leurs objectifs et de la nature des données à analyser.

Notre démarche n'est pas de choisir un modèle théorique et trouver une application qui permette de le valider. Au contraire, à partir d'objectifs précis et avec une application bien délimitée possédant certaines difficultés de représentation, nous voulons concevoir un modèle de résolution qui s'adapte au domaine. Un objectif est par exemple la **découverte** de règles pertinentes pour la classification en appliquant la méthode expérimentale fondée sur l'observation intime des faits. Ces règles n'auront pas forcément de signification statistique si l'on considère que la base d'exemples à traiter n'est pas *stochastique* mais bien *déterministe* [Mingers, 1987].

En effet, il est souvent difficile dans les applications en biologie d'émettre les hypothèses simplificatrices suivantes :

- ◊ complétude de l'ensemble d'apprentissage,
- ◊ tirage aléatoire des données,
- ◊ monotonie de la connaissance,
- ◊ nature de données (certaines ou probabilistes),
- ◊ existence d'une théorie du domaine complète et formalisée,
- ◊ indépendance des variables entre elles, etc..

Partant de ce constat, nous allons définir des critères d'appréciation de la robustesse qui englobent les facteurs qualitatifs précédents.

2.2.2 Les critères globaux d'appréciation

Dans notre approche de la robustesse, nous ne sommes définitivement pas dans un univers caractérisé par les probabilités et les lois *a priori*, mais bien dans un monde de diversité, d'incomplétude et où l'exception pourrait bien être la seule règle valide. Dans ce contexte, les critères d'appréciation de la robustesse seront les suivants :

2.2.2.1 Applicabilité à des domaines réels

Comme nous l'avons dit plus haut, le but est de résoudre un problème concret posé en biologie et de s'adapter au domaine étudié. Le but n'est pas de valider un modèle théorique déjà établi. Le rôle de l'informaticien est de suivre la démarche naturelle de l'expert. Ce critère fait appel aux facteurs de compréhension du domaine et d'adaptation aux besoins exprimés par les utilisateurs.

2.2.2.2 Un langage de représentation puissant

Le langage de représentation permet à l'expert de pouvoir exprimer toute sa connaissance dans les descriptions. Il ne doit pas être contraint par certaines limitations arbitraires de la logique (des propositions par exemple). Elles l'empêcheraient par exemple de représenter des objets de même type présents conjointement chez un même individu (voir la logique d'itération au § 4.4.6). De plus, cette représentation doit être transparente pour l'expert, la syntaxe de représentation ne le concernant pas, des outils conviviaux et interactifs d'aide à la description doivent l'assister dans cette tâche. L'exhaustivité et l'ergonomie sont les deux facteurs importants.

2.2.2.3 Facilité de mise en oeuvre par les utilisateurs

S'adapter au domaine, c'est prendre en compte les besoins des utilisateurs qui ne sont pas des informaticiens. Il faut donc leur fournir des outils suffisamment simples d'accès, attrayants et conviviaux. Il convient de faire attention à la complexité des paramètres d'apprentissage introduits que l'utilisateur final aura du mal à maîtriser. Ceux-ci peuvent être des choix de configuration d'interface, des choix de différentes mesures statistiques pour le traitement, des possibilités de pondération (coût de description d'un objet), des contraintes, des seuils...

En fait, il est nécessaire de bien observer les attitudes et comportements des utilisateurs finaux du système afin de leur fournir des outils dédiés à leurs besoins. Il faut se prémunir contre la tendance naturelle des informaticiens à vouloir fabriquer des outils "génériques" applicables dans n'importe quel domaine et que l'on appelle ensuite des "usines à gaz" du fait de leur difficulté de mise en œuvre et de leur inadéquation au problème posé. Chaque domaine possède sa propre spécificité à laquelle l'outil devra s'adapter s'il veut réellement répondre à une attente.

Inversement, le système ne doit pas être trop "spécifique" afin de ne pas devoir développer un nouvel algorithme à chaque fois que l'on change d'application. L'ergonomie, l'adaptation aux besoins exprimés et la compréhension sont les trois facteurs qui facilitent l'utilisation du système.

2.2.2.4 Incrémentalité

C'est une caractéristique fondamentale trop souvent négligée par les concepteurs car dépendant du choix des algorithmes retenus dans le système. Il existe deux définitions de l'incrémentalité dans la littérature. La première, dite **spatiale**, vise à traiter des bases de données de taille importante dans lesquelles il existe déjà une structure d'arbre de décision A. Sans qu'il soit nécessaire de reconstruire un arbre complet T à partir de tous les cas, ID5R [Utgoff, 1989] prend en compte l'ancien arbre A et à partir des nouveaux exemples, modifie la structure de manière à obtenir le même T. La seconde définition, dite **temporelle**, part du principe que l'apprentissage est un processus continu et donc les connaissances doivent évoluer à chaque fois qu'un nouvel ensemble d'expériences est réalisé.

En conséquence, le traitement des exemples doit s'effectuer par lot et les connaissances produites à partir des ensembles précédents sont modifiées pour prendre en considération les nouveaux exemples. Cette manière de procéder permet de pallier l'incomplétude de l'ensemble des exemples de départ. La prise en compte de ces deux définitions dans la conception du système rend possible son application sur des bases de données importantes et/ou incomplètes.

L'incrémentalité temporelle est celle que nous souhaitons appliquer. Elle fait appel aux facteurs de mise à jour, de cohérence (entre le modèle descriptif et les données) et de redondance (nouvelles données par rapport aux anciennes).

2.3 Discussion

Généralement, on dit qu'un système d'apprentissage est "robuste" s'il permet d'obtenir des résultats satisfaisants par rapport à un ensemble d'hypothèses de départ. Cette définition très générale de la robustesse laisse la porte ouverte à de multiples interprétations. L'appréciation du résultat est laissée au jugement de n'importe quel type d'utilisateur, qu'il soit informaticien, statisticien, biologiste, expert ou béotien. Or, les idées de ces différents utilisateurs sur la question ne sont pas toujours partagées, loin s'en faut !

2.3.1 L'informaticien

Pour lui, un système robuste traite des données pour obtenir des *résultats aussi bons que ceux de l'expert*. S'il possède une formation académique basée sur les mathématiques, il supposera que les exemples ont été recueillis convenablement selon un protocole d'échantillonnage précis. S'il est chercheur, le traitement est alors considéré comme la partie "noble" de l'acquisition des connaissances pour valider des solutions algorithmiques, parce que la phase de saisie des données est peu valorisable du point de vue scientifique. Il est d'ailleurs révélateur de constater que de son point de vue, le terme de validation des connaissances est dépendant du traitement qui a été préalablement effectué sur les données. Nous verrons dans notre approche que ce terme s'applique bien avant dans la phase d'acquisition des exemples à l'aide d'un questionnaire (la phase d'observation et de description est la véritable phase d'apprentissage pour le biologiste).

2.3.2 Le statisticien

Il argumenterait qu'un système robuste est doué d'une *forte capacité de prédiction sur des exemples qu'il n'a pas vus*, ce qui est le critère de qualité d'un bon système d'apprentissage. Il présuppose que les exemples à apprendre sont distribués selon une loi normale et correspondent à un modèle mathématique établi à partir des hypothèses suivantes :

◊ équi-possibilité des valeurs de chaque variable,

- ◊ indépendance des événements liés aux attributs (variables explicatives),
- ◊ tirage au hasard des individus de l'échantillon étudié (échantillonnage aléatoire) dans une population nombreuse et déterminée.

Certains statisticiens ont néanmoins une vision plus dynamique de la robustesse : pour Tomassone (1991), la Statistique est un guide pour toute démarche scientifique expérimentale. Elle demande de s'impliquer totalement dans l'analyse d'un monde “réel” incomplet et flou pour aboutir à sa représentation “virtuelle” obligatoirement schématique. Appliquer la Statistique requiert un assemblage *ad hoc* de trois composants : un Objectif, un Modèle, des Données.

L'objectif O correspond à un ensemble de questions auxquelles la Statistique est capable de répondre. Pour O fixé (ex : classification de plusieurs populations en classes homogènes), il existe au moins un modèle M qui permet de l'atteindre. Par modèle M, il faut comprendre deux éléments : une technique de sélection des données D (échantillonnage, plan d'expérience) et une technique de traitement des données quand on les aura acquises. Simultanément, un modèle M a besoin de certaines données D pour pouvoir être appliqué. Pour un utilisateur, il est indispensable de savoir quelles données D sont nécessaires pour utiliser M, et donc pour atteindre l'objectif O qu'il s'est fixé.

Pour ces statisticiens, la robustesse liée à l'acceptation du résultat découle d'un *va-et-vient entre M et D autour de O fixé* (figure 2.1). Cette robustesse est provisoire tant que des éléments nouveaux ne viennent pas contredire le résultat précédemment acquis.

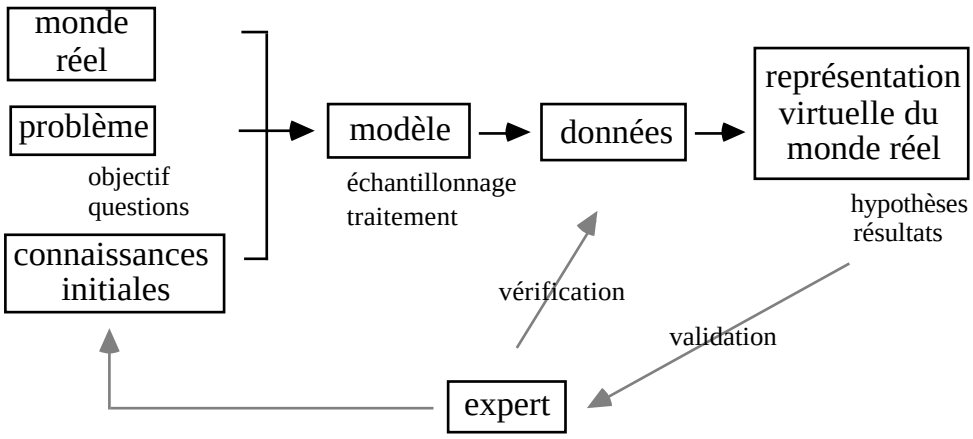

Fig. 2.1 : La robustesse dans la démarche statistique [Tomassone, 1991]

2.3.3 Le biologiste

Sa démarche est basée sur l'expérimentation. Conscient des problèmes liés à l'acquisition des connaissances sur du matériel vivant, il pourra dire qu'un système robuste est capable de *minimiser les erreurs dues aux "bruits"* dans l'acquisition des exemples. Lors du projet INSTIL, en tant qu'étudiants en agronomie, nous avons pu détecter différentes sources de bruits lors des phases de collecte, d'observation et de description des échantillons de plants de tomate malades. La figure 2.2 en donne une classification pratique :

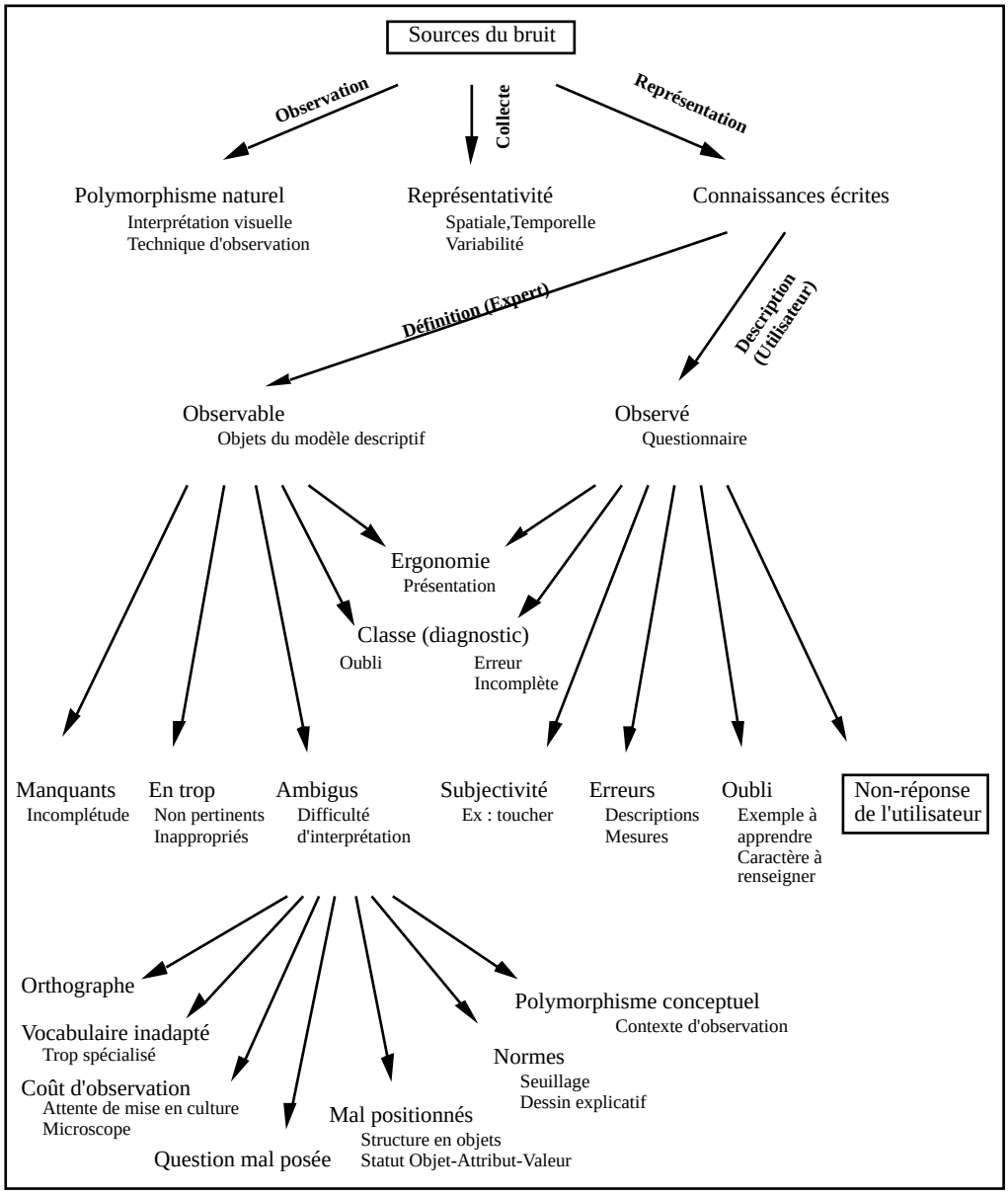

Fig. 2.2 : Classification des différents types de bruits dans INSTIL

L'un de ces bruits est la non-réponse de l'utilisateur à une question posée par le système expert lors de la procédure de détermination (en fait, la réponse est

«inconnu», ce qui n'apporte aucune information). Par exemple, le technicien agricole vient consulter le système de diagnostic TOM avec uniquement les fruits sur lesquels il observe des taches. Si l'arbre de décision engendré par KATE a choisi un premier critère de discrimination sur le feuillage (avez-vous observé des taches sur feuilles ?) et que l'utilisateur n'a pu faire l'observation demandée, le diagnostic obtenu risque d'être incertain. Étant confrontés à ce problème lors du démarrage de cette thèse, le terme de robustesse est apparu à ce moment pour y faire face. Il nous fallait trouver une solution pour résoudre ce bruit dans les consultations. Nous illustrerons la robustesse face aux valeurs manquantes en phase de consultation sur l'application des éponges marines (voir chapitre 7). Notons que cette expression a été utilisée par d'autres chercheurs en psychologie cognitive pour illustrer le même problème [Sutcliffe, 1986].

2.3.4 Le béotien

Il considérera le système robuste s'il "résiste" aux inexactitudes lors des réponses au questionnaire et qu'il arrive à résoudre son problème correctement tout en lui fournissant quelques explications. C'est son degré de satisfaction qui détermine son appréciation. Lorsque l'utilisateur est "naïf par rapport à l'observation", c'est-à-dire qu'il ne connaît pas la démarche d'expertise et n'a pas forcément une bonne pratique d'observation, il sera séduit par les capacités à la fois pédagogiques et de vulgarisation du système, se considérant peut-être lui-même comme un «bruit» pour le bon déroulement du raisonnement du système expert.

2.3.5 L'expert

Il auto-référencera la robustesse du système à sa propre manière "intuitive" de traiter les exemples. C'est la validité des conclusions du système qu'il est en mesure d'évaluer. Il s'agit là de son évaluation subjective sur la qualité d'une classification. Sa satisfaction peut être liée à différents facteurs [Niquil, 1993] :

- ◊ exactitude des règles apprises par rapport aux exemples soumis,
- ◊ présence ou absence souhaitée a priori de certains critères classificatoires,
- ◊ ordre de ces critères dans l'arbre de classification,
- ◊ degré de généralisation, etc..

Pour nous, l'objectif principal pour acquérir un système robuste est d'arriver à faire plus participer l'expert dans le fonctionnement du système car il est le garant de cette robustesse. Généralement, son rôle se borne à la fourniture de l'ensemble des exemples et à la validation des connaissances apprises. Il est effectivement intéressant de le faire intervenir au cours du traitement des exemples pour ajuster des paramètres et modifier le comportement du système.

Mais cela ne suffit pas. Le fonctionnement du système ne peut pas se réduire au simple traitement des données comme s'il s'agissait d'un aboutissement ! Comme pour l'approche statistique [Tomassone, 1991], nous avons bien conscience que l'acquisition des connaissances n'est pas un processus linéaire mais bien itératif et que le traitement n'est qu'un aspect (très marginal au niveau du temps consacré pour l'apprentissage) du fonctionnement global du système. Ce qui est aussi très important, c'est ce qui se passe avant et après le traitement des données afin de mieux maîtriser les variables et les exemples appris.

C'est pourquoi nous voulons aller plus loin dans cette thèse dans la formalisation des données **en amont du traitement** par les logiciels d'apprentissage automatique. Comme l'indique la figure 2.3, nous allons **expliciter** les connaissances initiales de l'expert au sein d'un modèle de l'observable. Les données observées devront s'y conformer, ce qui permettra d'obtenir des descriptions structurées comparables entres elles et d'atteindre l'autre objectif principal : **la robustesse des descriptions**.

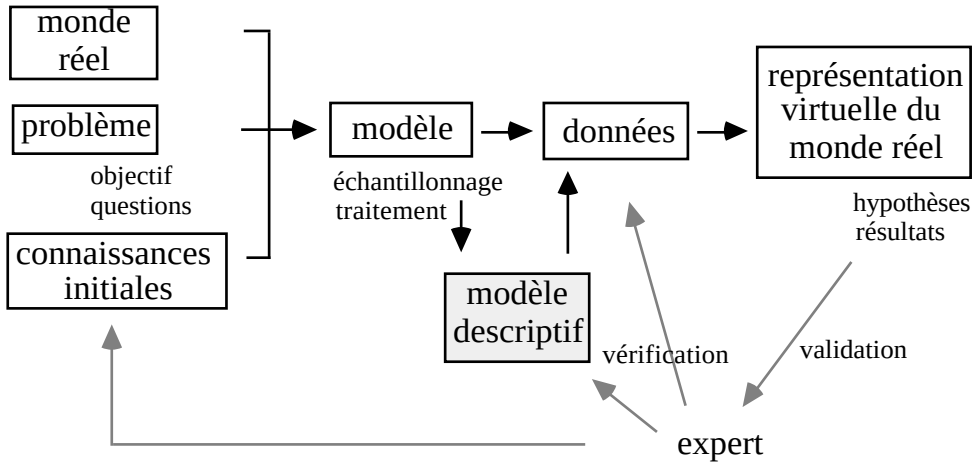

Fig. 2.3 : Comparaison de notre travail avec l'approche statistique

Nous nous apercevons donc que la robustesse est une notion toute relative, à manier avec une certaine précaution en fonction des interlocuteurs. Nous ne prétendons donc pas dans cette thèse fabriquer un système robuste de classification et de détermination des objets biologiques : cela est utopique dans un tel domaine. Nous souhaitons simplement apporter une contribution originale à son amélioration. Nous dériverons donc la robustesse au niveau de l'aide apportée par des outils informatiques, conçus de telle manière que l'utilisateur atteigne les objectifs qu'il s'est fixés (classification et/ou détermination) et maîtrise ainsi mieux son sujet d'étude. Il s'agit pour lui d'apprendre des choses nouvelles et utiles par ces outils, mais tout aussi bien sur son domaine que sur sa propre méthode de travail, ce qui contribuera à l'amélioration des connaissances générales.

2.4 Notre méthode d'acquisition des connaissances

L'amélioration de la robustesse passe par la mise au point d'une méthode d'acquisition de connaissances fondée sur l'observation des faits, calquée sur la pratique des biologistes systématiciens. La méthode est en conformité avec la démarche de tout scientifique utilisant le raisonnement “plausible” (l'induction) et l'analogie (le raisonnement par cas) à des fins de classification et détermination d’objets naturels.

2.4.1 Différents types de connaissances à acquérir

2.4.1.1 Connaissances de base (“background knowledge”)

Ce sont les connaissances des **faits observables** du domaine, exprimées dans le **modèle descriptif**. Elles recensent les objets observables liés entre eux par des relations, ainsi que leurs caractères observables (caractéristiques, propriétés, variables ou attributs) et les différents états possibles de ces caractères (valeurs ou modalités d'attributs). Ces objets permettent de décrire complètement une entité du domaine. Cette étape correspond à l'acquisition du modèle descriptif (phase 1).

Le générateur de modèles descriptifs est l'outil interactif qui permet de créer, d'éditer et de visualiser les objets graphiquement sous la forme d'un arbre. Cet outil de modélisation de l’observable est un composant d’HyperQuest™ (voir le chapitre 6, § 6.3). L'acteur principal de cette étape est l'expert du domaine assisté ou non du cogniticien.

2.4.1.2 Connaissances de faits observés

Ce sont des descriptions individuelles issues du remplissage d'un questionnaire hypertexte qui lui-même a été engendré automatiquement à partir du modèle descriptif (phase 2). Ces faits constituent les données en entrée du système d'induction ou de raisonnement par cas (phase 3).

Deux types de faits observés sont à considérer selon les objectifs du traitement :

◊ Classification : Le **cas** (ou exemple) est l'association d’une description d’objets et de l’identification de la classe à laquelle appartient l'individu possédant ces objets. La constitution d'une base de cas permet d’atteindre

la caractérisation (définition) des différentes classes d'affectation prédéfinies, et par suite un système expert de détermination.

◊ Détermination : L'**observation** est une description d'objets sans classement associé à l'individu (le nom de la classe). Une observation permet de consulter le système pour déterminer l'individu.

Ici, le nom de la classe est une sortie, alors que c'est une entrée dans le cas de la classification.

L'outil qui permet de créer, d'éditer et de visualiser les cas et les observations s'appelle le **questionnaire**. Il est construit automatiquement à partir du module générateur de questionnaire interactif hypertexte de l'outil HyperQuest™ (voir chapitre 6, § 6.4). Cet outil exploite les connaissances du modèle descriptif de manière à les présenter simplement sous forme de cartes et de naviguer entre elles pour renseigner les différents objets. Ce questionnaire est personnalisable et permet d'intégrer des images pour illustrer les objets à décrire. Il permet aussi de rappeler des cas pour les éditer. Le questionnaire produit en sortie une description qui est soit un cas alimentant la base de cas pour la classification, soit une observation brute à soumettre au système pour identification.

Les acteurs de cette phase sont aussi bien l'expert du domaine que les futurs utilisateurs du système de détermination.

2.4.1.3 Connaissances produites

◊ Classification : ce sont les **règles** ou **l'arbre de décision** induits automatiquement par généralisation des exemples et traduisant une conjonction de propriétés à satisfaire pour appartenir à la classe nommée en conclusion de la règle ou à la feuille de l'arbre (phase 3),

◊ Détermination : la connaissance recherchée est **l'identification** de la classe d'appartenance de l'individu à déterminer.

Toutes ces connaissances sont différentes par nature et s'acquièrent dans cet ordre prédéterminé afin de fabriquer le système de classification et de détermination. Cela signifie que ces étapes doivent s'enchaîner en commençant par la phase 1 qui est indépendante des deux autres. La phase 2 dépend de l'élaboration de la phase 1 et la phase 3 dépend à la fois des deux autres (de la sorte, les trois phases ont en commun le modèle descriptif).

2.4.2 Deux types de traitements des exemples pour la classification et la détermination

En fonction des deux objectifs de l'expérimentation (classification ou détermination), deux types de traitement des exemples sont proposés.

Pour la **classification**, qui concerne surtout l'expert, une caractérisation des classes peut être obtenue par généralisation des exemples (apprentissage) et présentée sous forme d'un arbre de décision. L'expert peut être amené à tester l'incidence de différents critères de généralisation (pondération, efficacité, coût) en comparant les différents arbres ainsi engendrés. La technologie de l'**induction** nous paraît la mieux adaptée à l'objectif de classification.

Pour la **détermination**, qui concerne aussi bien l'expert que le biologiste, l'objectif est d'extraire progressivement de la base d'exemples ceux qui ne sont pas en contradiction avec la nouvelle observation à déterminer, jusqu'à se confiner dans une classe. La technologie du **raisonnement par cas** nous semble préférable à l'induction en ce qui concerne l'objectif d'identification (voir le chapitre 7).

On a représenté sur le schéma de la figure 2.4 ci-dessous les trois étapes permettant d'acquérir les connaissances descriptives ainsi que les deux types de traitement utilisés :

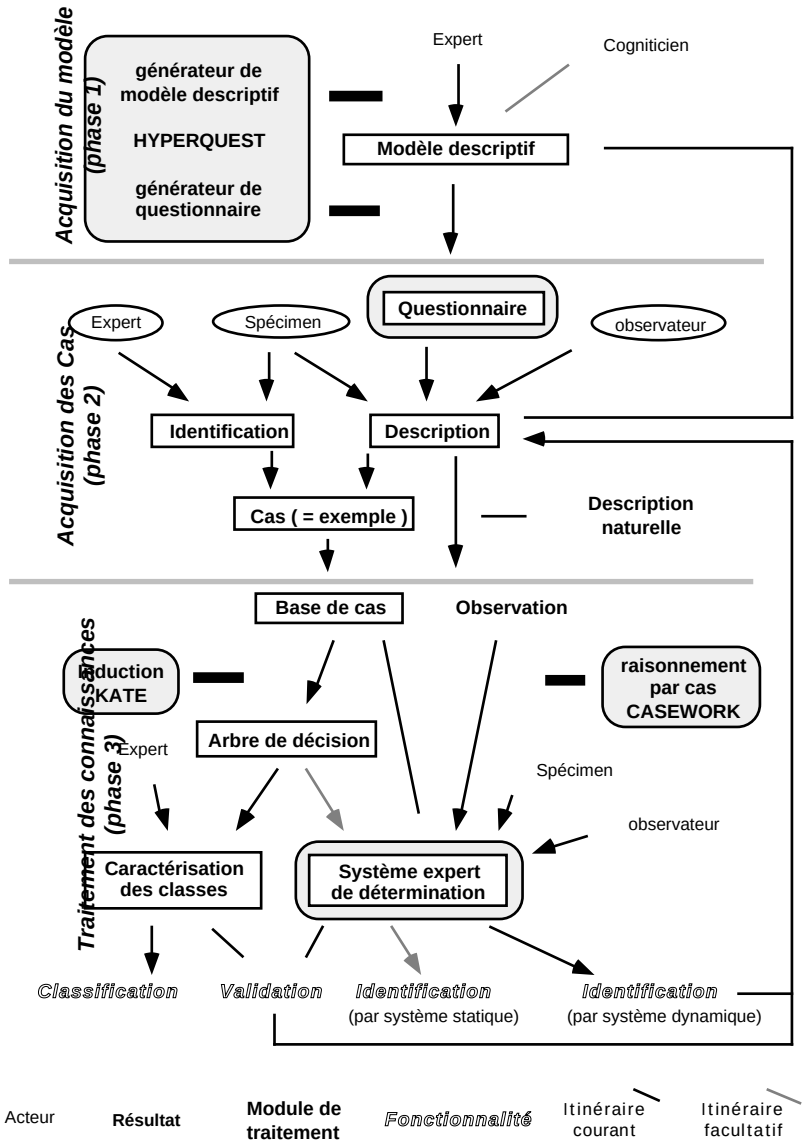

Fig. 2.4 : Synoptique de notre méthode d'acquisition des connaissances

Deux types de détermination sont possibles avec le système actuel :

❶ Le premier utilise un système statique de détermination. **KATE** fabrique un arbre de décision ou des règles de décision, ce qui forme une généralisation des cas dont on ne retient que les critères de détermination les plus efficaces (système figé et maximalement discriminant).

❷ Le second est un système expert dynamique de détermination. **CaseWork** raisonne directement à partir des cas en appliquant un principe d'analogie pour retrouver ceux les plus semblables à l'observation courante (système dynamique).

En effet, la phase de consultation du système met en jeu toutes les connaissances décrites préalablement et permet la détermination d'un nouveau cas. Elle intervient après que la phase d'induction par KATE ait engendré un arbre de décision pour le système statique alors que pour le système dynamique, les deux phases sont imbriquées au cours de la détermination : l'utilisateur guide la discrimination en fonction des réponses qu'il donne (ou ne peut pas donner) au cours de la consultation.

Le questionnaire du domaine peut intervenir en phase de consultation pour le système dynamique. Il permet à l'utilisateur de ne pas être obligé de suivre un chemin de l'arbre de décision avec les questions posées relatives à chaque nœud. L'utilisateur fournit dans un premier temps sa propre description de ce qu'il observe avec le questionnaire, puis cette description est interprétée par CaseWork qui pose des questions complémentaires s'il n'aboutit pas à un résultat certain. Le questionnaire favorise donc la maîtrise par l'utilisateur de la consultation du système de détermination dynamique.

De plus, si cette consultation peut être validée par l'expert, la description issue du questionnaire et la détermination de l'expert constituent un nouveau cas qui peut être introduit dans la base initiale.

2.4.3 La phase de validation des connaissances apprises

Une dernière phase mérite d'être mentionnée : il s'agit de la **validation**. Elle permet de détecter des incohérences, que ce soit avec l'arbre de décision de KATE ou bien avec le système de détermination de CaseWork. Quand un résultat s'avère invalide, trois causes peuvent être invoquées :

1 - Une description a été mal renseignée (valeur erronée par exemple),
2 - La base de cas est incomplète, non représentative de la variété réelle,
3 - Le modèle descriptif est incomplet (critère discriminant oublié).

Les deux premiers problèmes sont ponctuels et peuvent être résolus simplement en retrouvant la description erronée ou en rajoutant un exemple. Le dernier problème relève de la structure même du modèle descriptif et a pour conséquence la remise en question non seulement de la base de cas mais encore la mise à jour du questionnaire afin d'assurer sa cohérence avec le modèle descriptif.

Dans la méthode mise au point, toute la chaîne des outils allant de la construction du modèle descriptif au traitement des exemples par induction et raisonnement par cas est complète. Néanmoins, la phase d'itération sur le modèle à modifier n'est pas prise en compte au niveau des anciens cas. En effet, si un nouveau questionnaire est généré prenant en compte de nouveaux objets, attributs ou valeurs possibles, la modification de la structure du modèle ne remet pas à jour

l’ancienne base de cas. Ceux-ci doivent être complétés manuellement avec un traitement de texte pour être conformes au nouveau modèle.